METHODS FOR ENRICHING TRANSPORTATION SURVEY DATASETS: WITH SAMPLE APPLICATIONS USING PSYCHOMETRIC VARIABLES

A Dissertation
Presented to
The Academic Faculty

by

Faaiqa Atiyya Shaw

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Civil and Environmental Engineering

Georgia Institute of Technology
May 2021

METHODS FOR ENRICHING TRANSPORTATION SURVEY DATASETS: WITH SAMPLE APPLICATIONS USING PSYCHOMETRIC VARIABLES

Approved by:

Dr. Patricia L. Mokhtarian Advisor
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Giovanni Circella
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Michael P. Hunter
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Kari E. Watkins
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. James S. Roberts
School of Psychology
Georgia Institute of Technology

Dr. Ram Pendyala
School of Sustainable Engineering and the Built Environment
Arizona State University

Date Approved: January 27, 2021

بسم الله

ACKNOWLEDGEMENTS

I am indebted to my faith alongside my strong network of family, friends, colleagues, mentors, and advisors who made this journey, and the resultant growth possible. Special thanks to my thesis defense committee – Dr. Mokhtarian, Dr. Hunter, Dr. Roberts, Dr. Watkins, Dr. Circella, and Dr. Pendyala for your guidance, comments, and support throughout this process.

Advisors and mentors

To my core advisors, Dr. Michael Hunter, Dr. Patricia Mokhtarian, and Dr. James Roberts – you welcomed me into your guidance, believed in me (and pushed me) without reservation, and supported my dreams and endeavors as if they were your own. You were the ones I laughed with most. You were the ones I suffered with most. You have become family. Thank you.

To the Georgia Tech transportation faculty who treated me as if I were their own from day one, Dr. Rodgers, Dr. Watkins, Dr. Pendyala, Dr. Circella, Dr. Guin, Dr. (Ann) Xu, Dr. Kennedy, Dr. Gbologah – you believed in me with the same tenacity as my advisors, personally mentoring and guiding me at different points throughout this journey with kindness and friendship. Thank you.

To the Georgia Tech administration, faculty, and staff with whom I have had the privilege of developing deep friendships and/or receiving valuable mentorship from over the years, Dr. Reggie DesRoches, Dr. Larry Jacobs, Dr. Robert Simon, Joshua Stewart,

Jess Hunt-Ralston, Marjorie Jorgensen, LaJauna Ellis – just a small sample of the people that made Georgia Tech the nurturing and supportive home it was for me. Thank you.

Family and friends

To my husband, Ali Etezady, whom I met on this journey, you have been and continue to be the single greatest driver of my personal growth while supporting my professional growth in every way possible. You see me at my worst and still continue to believe only the best. Thank you. I love you.

To my parents, your sacrifices for me, Naimah, and Fawwaz are more noticed and appreciated than you realize. We can never repay you, but I hope that we have made you and will continue to make you proud, both through our impact on our families as well as on the world. My love and prayers for you are immeasurable.

To my siblings and their partners, Fawwaz – my commitment to matching your intelligence from a young age (irrespective of our nine-year age difference) is at least partly responsible for my educational success and drive. Naimah – your commitment to continuous character development makes me proud to call you my role model in both personal and professional aspects of life. Zaheir and Jayme – you have become integral members of our family, and two of my closest and most supportive friends – I feel truly blessed to have gained you as my additional brother and sister. Family is forever.

To my nieces and nephews, Za'id, Hadiyah, Saeed, Rayyan, Hanna, Zahra, Zain, Sara, Aleena – you were my greatest source of joy during my long higher education journey over the last 10 years. I love every single one of you with a special, unique, and unwavering

devotion – and I hope that my journey will be inspiring to you as you craft your own paths. I am your biggest supporter, always.

To my dear friends and extended family – I am truly blessed that the list is too long to specify, but you know who you are – you provided me conversations that revolved around things other than transportation, you grounded me, you inspired me, you motivated me, you comforted me. Thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

ACS	American Community Survey
BLM	Binary Logit Model
CCPA	California Consumer Privacy Act
CFA	Confirmatory Factor Analysis
CNT	Center for Neighborhood Technology
CV	Common Variable(s)
DRL	Deterministic Record Linkage
EFA	Exploratory Factor Analysis
EPA SLD	Environmental Protection Agency Smart Location Database
EU	European Union
GDOT	Georgia Department of Transportation
GDPR	General Data Protection Regulation
GTFS	General Transit Feed Specification
HHTS	Household Travel Survey
ICT	Information and Communication Technologies
kNN	k-Nearest Neighbor
LCCM	Latent Class Choice Model
LU	Land Use
MAPS	Microscale Audit of Pedestrian Streetscapes
MI	Multiple Imputation
ML	Machine Learning
MPO	Metropolitan Planning Organization
NHTS	National Household Travel Survey

PCA	Principal Components Analysis
PRL	Probabilistic Record Linkage
RF	Random Forest
SED	Socioeconomic and Demographic
SVM	Support Vector Machine
TM	Targeted Marketing
TUS	(United Kingdom) Time Use Survey
XGB	Extreme Gradient Boosting

SUMMARY

Technological disruptions, environmental and health upheavals, and societal shifts are just a few of the major forces interacting in rapid and unprecedented ways to influence how we live, work, and navigate within our built environments. Transportation engineers and urban planners must grapple with how such widespread, and in many cases, yet unknown, changes will alter the urban landscape, shifting travel patterns and requiring a fresh look at infrastructure forecasting, planning, and development into the future. In a time of such uncertainty, it is increasingly important for national, state, and regional planning organizations to be able to understand and forecast behavioral and attitudinal changes. However, modeling such shifts depends on actively collected survey data, which are infrequently gathered, time and cost-intensive, and suffer from continuously declining response rates (and accompanying biases).

Accordingly, the work presented in this thesis aims to address some of these challenges by making use of data driven tools like machine learning, and psychometric-based approaches like latent variable analyses, within the context of the rapidly growing big data landscape to develop and present three approaches for supplementing and/or expanding transportation survey datasets using active and passive data streams. Broadly, these approaches include: (1) exploring and utilizing new sources of data for transportation modeling and analysis; (2) integrating and expanding existing, traditional sources of transportation data with both active and passive data sources; and (3) developing marker statements for expanding the breadth of information obtained without substantially increasing survey lengths. These methods are demonstrated by applying them to enrich

traditional transportation surveys with psychometric data (e.g., attitudes), which have been shown in the literature to have the ability to explain and predict behaviors, but which are often not captured on surveys.

Each application is validated by integrating the enriched datasets within sample travel behavior models, and observing changes in predictive accuracy, model fit, and interpretability. Findings show that expanded variable richness for transport surveys, specifically with psychometric variables like attitudes, can improve performance and interpretability of travel behavior models. This research has societal implications that center on the potential for improved travel demand forecasting and behavioral predictions. Such improvements can facilitate more efficient expenditures, improve infrastructure planning, and ultimately increase quality of life for all. Even more broadly, the methods of this research may be applied to enrich many more large-scale behavior-based surveys with diverse variables, thereby providing richer, more robust data streams for use in an array of modeling and forecasting efforts.

CHAPTER 1. INTRODUCTION

1.1 Background and motivation

We are poised on the cusp of transport disruption that may only be likened to that period of time in the late 19th and early 20th centuries when Karl Benz and Henry Ford changed the way people moved forever. Already, we are seeing widespread changes in the form of technosocial trends such as: (a) the ever-expanding reach and capability of mobile information and communication technologies (ICTs); together with (b) the increasing "passengerization" of travel – exemplified by the onset of vehicular automation, coupled with growing (albeit demographically and geographically uneven) market penetration of ridesharing and transit. In addition to technology-driven changes, recent events have underscored the potential impacts of environmental (e.g., climate change related disasters) and health upheavals (global pandemic) on the transport system. As transportation professionals, we must grapple with how such widespread, and in many cases, yet unknown, changes will alter the urban landscape, shifting behavioral patterns and requiring a fresh look at infrastructure forecasting, planning, and development into the future. In a time of such uncertainty, it is increasingly important for national, state, and regional planning organizations to be able to understand and forecast transport-related shifts in behaviors.

However, the inaccuracies of transport forecasting models are well documented (Bain, 2009; Hartgen, 2013; Nicolaisen & Driscoll, 2014; Parthasarathi & Levinson, 2010; Voulgaris, 2019; Welde & Odeck, 2011), with current models often operating at less than 10% explanatory power, and requiring subjective alterations with derived parameters to

improve ("match existing conditions") performance. While the difficulties of forecasting travel choices and patterns are not surprising given the inherent challenges associated with predicting human behavior, there is consensus among transportation professionals that more could be done to improve model performance. In transportation, improving transport model performance has generally been pursued through two primary avenues: (1) improving data quality and richness (Welch & Widita, 2019); and (2) increasing model complexity and/or using data-driven approaches such as machine learning algorithms (Cheng, Chen, De Vos, Lai, & Witlox, 2019; Zhao, Yan, Yu, & Van Hentenryck, 2020). Research in the latter domain dominates the literature, which is intuitive given that the former approach is often constrained by data availability, resource limitations, or other such challenges that may be outside the control of the analyst.

Research in the data improvement domain has primarily centered on the rapidly proliferating big data landscape, which has created fertile ground for the exploration of novel data sources to support transportation supply and demand modeling applications; however, forecasting travel behavior still primarily depends on household and individual-level survey data. The proposed thesis contributes to the literature in this domain in two distinct ways: (1) by distilling methods/frameworks for enriching transport survey datasets using *both* novel and existing data streams, and (2) by applying and validating the developed methods on travel behavior models. Surveys are expected to remain the key source of data for travel demand forecasting and behavioral models in the foreseeable future due largely to the user-verified, self-reported nature of survey responses, alongside their ability to obtain domain-specific data that often is not (easily) available through other data streams. However, the advantages of survey data are tempered by several critical

challenges, key among them: (1) the infrequency of survey data collection efforts, a result of the resource-intensive nature of survey development, dissemination, and post-processing efforts; (2) continuously declining survey response rates that can threaten the validity of survey findings by increasing non-response bias; and (3) a lack of breadth/variable richness in domain-specific survey instruments, a partial byproduct of increased nonresponse rates attributed to longer surveys. Although there is currently an array of approaches for addressing survey-related challenges – for example: larger, varied incentives, mixed sampling methods, and increasingly complex approaches for redistributing survey responses to better represent target-area population distributions (i.e., weighting) – researchers continue to face heightened concerns regarding growing survey *non-response rates*, a key culprit in the challenge to obtain high quality, long-form, representative survey data.

To address some of the afore-mentioned shortcomings of survey data, this thesis details and applies three approaches for supplementing and/or expanding transportation survey datasets using active and passive data streams. These methods are demonstrated by applying them to supplement and enrich transportation surveys with psychometric data, which have been shown in the literature to have the ability to explain and predict behaviors (Domarchi, Tudela, & González, 2008; Kuppam, Pendyala, & Rahman, 1999; Mokhtarian & Salomon, 1997). In addition to empirical support from the literature, it is also conceptually clear that a model that is able to capture the influence of pro-environmental attitudes/values, or attitudes towards privacy, safety, and the sharing economy, may be better able to predict environment-related behavioral changes in vehicle-miles traveled, or receptiveness to ridesharing transport options, respectively. Over time, the efforts

initialized by this work are intended to provide more diverse and robust data streams for use in transportation supply and demand models, thereby potentially improving current urban and regional forecasting models.

1.2 Research objectives and contributions

To advance transport forecasting and behavioral modeling, this thesis presents three approaches for expanding transportation survey datasets, the key data sources needed to better understand the choices that individuals make within the context of the built environment. The survey data enrichment methods detailed in this document include:

1. utilizing and integrating novel sources of data;
2. expanding transportation survey datasets through predictive transfer; and
3. abbreviating survey instruments/questionnaires using marker statements.

The presentation of each method encompasses a conceptual framework and/or a grounding of the method within the transport literature, as well as a complete application of the method from data integration to validation.

The *applications* of the presented methodological approaches focus on the enrichment of transportation datasets with psychometric variables, which are individual-specific variables such as attitudes, preferences, perceptions, social and personal values, and other such user traits. These types of variables are selected for demonstrating the utility of the methods because a lack of psychometric traits available for use in forecasting and behavioral models has been identified in the literature as a contributory factor to poor transport model performance. As before noted, in this era of rapid introduction of disruptive

technologies and services, improving the flexibility and realism of transport models has never been more imperative. In addition, the lack of psychometric variables available for use in transport modeling is directly related to survey-related challenges such as the inherent difficulties (e.g., reduced response rate) associated with longer surveys, as well as the more nuanced survey design needed for obtaining individuals' psychometric traits.

To summarize, this work aims to make contributions in both methodological and applied domains, detailing methods that transport analysts can use to expand their datasets, and then applying these methods to bring psychometric variables into survey datasets and ultimately, travel behavior models. Figure 1.1 visually summarizes the work proposed in this thesis, from motivation and problem definition to methodological approaches and application.

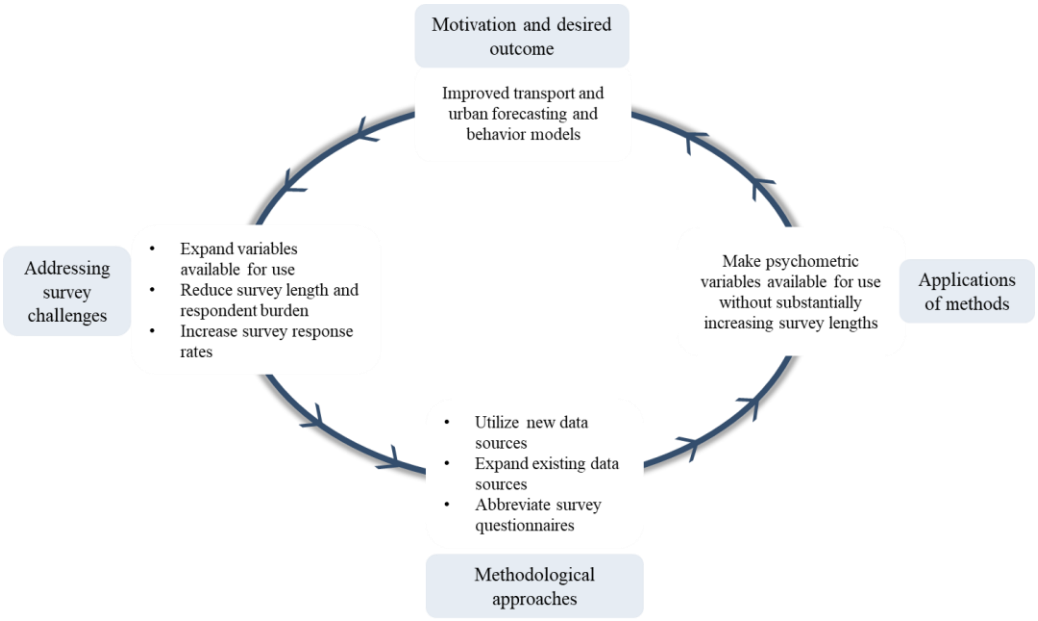

Figure 1.1. Overview of study, from motivation to desired outcome

1.3 Overview of survey data in thesis

In this thesis, two main survey instruments are utilized for the applications of the methods: (1) a statewide transportation survey conducted (by our research team at Georgia Tech) for the Georgia Department of Transportation (GDOT survey) and (2) the Georgia subsample of the U.S. National Household Travel Survey (NHTS), a nationwide travel behavior-focused survey conducted by the U.S. Department of Transportation. Here, summaries of the survey datasets are provided, and further information is given as needed throughout the document.

1.3.1 Georgia Department of Transportation Survey (GDOT survey)

The Georgia Department of Transportation Emerging Technologies Survey (GDOT survey; conducted September 2017 to January 2018) is a statewide research-oriented transportation survey that obtained general attitudes and preferences, technology use, lifestyle-related variables such as employment and relationship status, a wide array of current and future travel-related attitudes, behaviors, and preferences, and socio-economic/demographic characteristics (Kim, Mokhtarian, & Circella, 2019). Invitations to complete the GDOT survey were mailed to two groups of respondents: (a) a randomized set of 30,000 names/addresses selected from across 14 Metropolitan Planning Organization (MPO) areas in Georgia (this randomized set of names/addresses was purchased in Fall 2017 from a TM data provider); and (b) ~5000 individuals who responded to the NHTS and agreed to be contacted for a follow up survey.

Approximately 1800 of the randomly sampled 30,000 respondents returned a completed (usable) GDOT survey (termed the GDOT_R subset in this thesis), and about

1500 of the ~5000 NHTS respondents sampled returned a usable GDOT survey (termed NHTS_Agree_R, for "Agreed to be contacted again, and Responded to the subsequent GDOT survey contact"). Thus, roughly 3300 valid respondents were retained in the GDOT dataset. See Figure 1.2 for a visual representation of the GDOT and NHTS sample subsets used in this thesis and see Table A1 in Appendix A.1 for descriptive statistics on the GDOT sample. For additional details on the survey, please see Kim et al. (2019).

1.3.2 National Household Travel Survey (NHTS)

The NHTS is a repeated cross-sectional travel survey conducted by the Federal Highway Administration, and deemed the "authoritative source on travel behavior of the American public" (Federal Highway Administration, 2018). The NHTS used in this study was the most recent wave, conducted from March 2016 to May 2017, and includes both individual and household-level modules that cover general household characteristics, vehicle ownership attributes, long distance travel behavior, and person-level characteristics including person trips (for a chosen travel day) and health. Additional details regarding the NHTS can be accessed at https://nhts.ornl.gov/documentation.

As mentioned previously, approximately 5000 respondents from the Georgia subsample of the NHTS agreed to be contacted again for a follow up survey, and these respondents received a GDOT survey several months after completing the NHTS. Of these, ~1500 usable returns (the ***NHTS_Agree_R*** subsample) represent respondents for whom *both* GDOT and NHTS data is present (i.e., an overlapped sample). The remainder of the 5000 respondents represent individuals who agreed to be contacted again, and thus received a copy of the GDOT survey, but did not respond to it (***NHTS_Agree_DNR***, Agreed but

Did Not Respond to the subsequent GDOT survey contact). Of the total NHTS Georgia subsample, ~3500 respondents indicated that they did not want to be contacted again, and as such did not provide shareable name and address information (***NHTS_DNAgree***, Did Not Agree to be contacted again for a follow-up survey). Thus, these three NHTS subsets along with the GDOT-only subset (***GDOT_R***) represent **four** distinct subsets of respondents that comprise the transportation survey datasets used in this thesis (see Figure 1.2 for a schematic depiction of the subsets and Table A1 in Appendix A for SED characteristics).

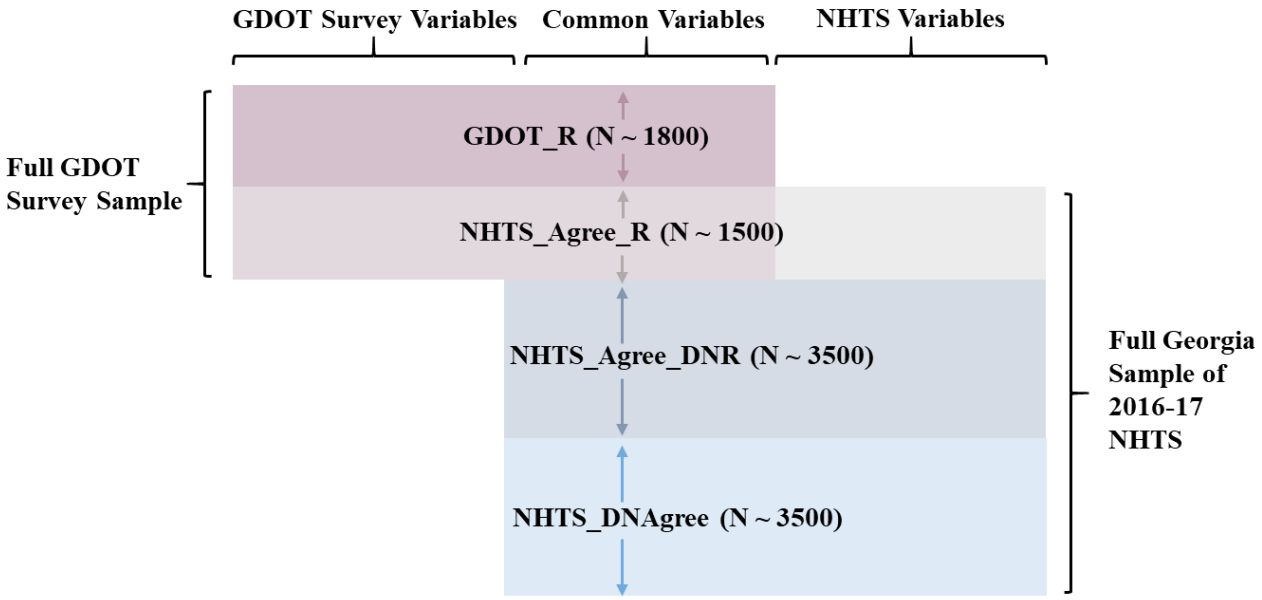

Figure 1.2. Schematic representation of GDOT and NHTS data subsets

CHAPTER 2. UTILIZING AND INTEGRATING NOVEL DATA SOURCES

For a data stream to have the opportunity to be a viable data source in transportation (or any field, for that matter), analysts must be able to: demonstrate that the data source has clear benefits, acknowledge the benefits or challenges of utilizing the data source, ensure that the source is accessible by other users, and most importantly, there must be replicated, independent examinations of the internal and external validity of the data within the relevant domain-area of application. In this chapter, targeted marketing (TM) data is investigated as a potential reusable, accessible novel source of data that can be used to supplement transport modeling efforts. Thus, the purpose of this chapter is two-fold: (1) to provide a model of the components that are deemed critical when presenting new data sources; and (2) to provide a discipline-specific resource that details potential applications and impacts of TM data in transportation.

Accordingly, this chapter provides:

1. a thoroughly researched background section defining TM data and its component data sources, and exhaustively examining the associated benefits and challenges of the data source within transportation;
2. a conceptual typology of potential TM data applications in research and practice, with examples pulled from the literature when applicable;
3. a data integration framework that encompasses data acquisition and preparation, informed by case studies integrating TM data with different types of transportation surveys/datasets; and

4. a series of internal validation exercises that examine TM data biases using varied methods.

In Chapter 5, TM data is externally validated by examining its performance in a range of travel behavior models. By undertaking a systematic approach to investigating TM as a transportation data source, the intent is to model an approach that can be used by analysts working on other nontraditional data streams.

The work detailed in this chapter is from the following manuscript which is currently under review:

Shaw, F. A., Wang, X., Mokhtarian, P. & Watkins, K. (paper under review, available upon request from authors). Supplementing transportation data sources with targeted marketing data: Applications, integration, and validation.

2.1 Abstract

Unlike many third-party data sources, targeted marketing (TM) data constitute holistic datasets, with disaggregate variables – ranging from socioeconomic and demographic characteristics to attitudes, propensities, and behaviors – available for most individuals in the population. These qualities, along with ease of accessibility and relatively low acquisition costs, make TM data an attractive source for the supplementation of traditional transportation survey data, which are facing growing threats to quality. This chapter develops a typology demonstrating ways in which TM data can aid in the design of transport studies, as well as in the augmentation of modeling efforts and policy scenarios, allowing for improved understanding and forecasting of travel-related attributes. However, challenges associated with integrating, validating, and understanding TM

variables have resulted in only a few transportation studies that have used these data thus far. In this chapter, a transportation discipline-specific resource for TM data is provided, informed by the integration of an extensive TM database with both the National Household Travel Survey (Georgia subset, NHTS) and a statewide travel behavior survey conducted in Georgia on behalf of the Georgia Department of Transportation (GDOT Survey). Using the resultant datasets, TM data is internally validated by means of several approaches; and it is found that the TM dataset reports gender, age, tenure, race, marital status, and household size with match rates ranging from 70% to 90% relative to both transportation surveys. However, biases are identified in favor of population segments that may have more longstanding financial/transactional records (e.g., males, homeowners, non-minorities, and older individuals), biases comparable but not identical to those of survey data. While this work suggests wide-ranging implications for the use of TM data in transportation, it is cautioned that flexible and responsible approaches to using these data are critical for staying abreast of evolving privacy regulations that govern third-party data sources such as these.

Keywords: consumer data; targeted marketing data; travel behavior; household travel survey; big data; third-party data; travel demand modeling

2.2 Introduction

Declining travel survey response rates coupled with the rapid proliferation of big data have created fertile ground for the exploration of novel third-party data sources to support transportation supply and demand modeling applications. Most prolific have been the use of mobile phone location data to supplement traditional travel diary data, but a wide range of sources, from social media to smart cards, have been effectively used to provide/augment key transport model inputs (Chen, Ma, Susilo, Liu, & Wang, 2016; He, Miller, & Scott, 2018; Khan, Ngo, Morris, Dey, & Zhou, 2017; Ma, Li, Yuan, & Bauer, 2013; Ruiz, Mars, Arroyo, & Serna, 2016; Toole et al., 2015; F. Wang & Chen, 2018; Z. Wang, He, & Leung, 2018; Welch & Widita, 2019). These successes make clear that transportation planners, engineers, and researchers must continue to explore effective approaches to utilizing nontraditional data sources in transport modeling and forecasting efforts. However, the sources utilized thus far have tended to entail siloed data that lack linkages to socioeconomic and demographic (SED) indicators, psychometric attributes (e.g., attitudes), and behaviors across different domains. In contrast, targeted marketing (TM) data are largely untapped, low-cost, holistic databases that house hundreds to thousands of diverse variables on individuals and households across the country.

TM data are typically used to identify and market to individuals likely to be more receptive to a particular product/brand, but due to attributes such as data magnitude and ever-increasing variable richness (supported by continuous technological advances), there is enormous potential in using these data to supplement travel demand modeling and forecasting efforts that currently primarily depend on actively collected survey data. Transportation surveys such as household travel surveys (HHTS) and research-oriented

stated and revealed preference surveys are infrequent, expensive, and suffer from continuously declining response rates that can threaten the validity of using these sources independently (PTV NuStats, 2011, National Research Council, 2013). On the other hand, while TM data are available and relatively inexpensive, challenges associated with integrating TM data with transportation survey data, validating acquired TM variables, and further interpreting these variables have meant that only a few transportation studies have successfully used these data.

The remainder of this chapter is organized as follows. The chapter begins by examining sources that inform the creation of TM data, and details some benefits and challenges associated with utilizing these data (Section 2.3). Next, a review is presented on how TM data have thus far been used in the transport domain, and a taxonomy developed of possible transportation applications, outcomes, and research directions that could benefit from the use of TM data (Section 2.4). Based on the integration of a large TM dataset with statewide and national transportation surveys (Section 2.5), a framework is then presented for the integration of TM data with existing transportation data sources (framework summarized in Section 2.5, and further detailed in Appendix B.2). Using the integrated dataset, the quality of TM data is examined relative to comparable self-reported data from travel surveys, and the biases of TM data are explored by comparing survey respondents with and without records in the TM database (Section 2.6). Next, recommendations for how transportation professionals can address identified TM data biases are provided (Section 2.7). The chapter closes with a summary of contributions and findings (Section 2.8). Appendix B.1 provides supplementary tables and figures, while Appendix B.2

provides additional data integration details for analysts seeking to enrich their own survey datasets with TM data.

2.3 Exploring targeted marketing data

Since TM data have been little-used by the transportation community as yet, the discussion begins with a general introduction to this type of data. In this section, TM-related terms and data sources are examined, followed by a summary of benefits and disadvantages of which transportation researchers/ practitioners should be aware when using TM data.

2.3.1 Defining targeted marketing data

The terms consumer, audience, and/or (targeted) marketing data are often used interchangeably; however, they can refer to different concepts. In this work, the term "Targeted Marketing (TM) data" is used; the reason for this distinction is explained by presenting a brief overview of the related terms here:

- **Consumer data** are defined by Birkin (2019) as "data arising from the interaction between customers and service providers", and should be the byproduct of a "market-based exchange of value". The most common form of consumer data is transactional data, which are obtained each time a consumer utilizes a credit or debit card to make a purchase. These data are then typically aggregated to yield variables such as the number of purchases made within various consumer categories (e.g., apparel, home, etc.), frequency of purchase, and the medium used for transactions (e.g., online, in-store, etc.). Consumer data may also include less traditional,

technologically-enabled transactional interactions such as mobile application use, digital browsing, and smart card usage for fare payment.

- **Targeted marketing data** refers to large databases that house hundreds to thousands of individual- and household-level variables that data providers (often these are credit reporting firms) either directly collect, purchase, or develop. TM data are developed with the explicit purpose of being re-sold to businesses who use selected variables to aid in marketing campaigns that target their specific audience. In some contexts, TM firms use the term "consumer data" to indicate that TM variables represent profiles of consumers in the marketplace. As such, the term "TM data" is often conflated with the term "consumer data"; however, while TM databases often include many variables that are derived from consumer data, they also include other types of variables/data.
- **Audience data** is a term used by marketers/business strategists to represent variables that are specific to a business's target base of consumers, i.e., its audience. Business entities may select from already developed audience segments present in TM databases, or alternately, may request TM providers to develop personalized segments that are relevant to their services. Thus, audience data/segments can be derived from TM databases, although businesses also often collect their own internal audience data.

As can be seen, there is significant overlap between these terms. It is recommended that analysts use the term "TM data" for datasets purchased from TM and/or credit reporting firms or other large third-party data providers/ compilers, as it is likely that many of the variables in such databases have been developed and/or imputed based on a host of

other variables. For example, while a variable description may suggest that a variable is a "pure" consumer variable (i.e., directly collected by a service provider), it is likely that this variable was modified using information from other sources (e.g., from public records or survey data) in the TM database, and thus the use of the term "TM data" aids in clarifying the source of the variables being used.

Figure 2.1 provides a non-exhaustive organizational structure for sources that typically inform TM databases. Shown first is the most established source, that of administrative data such as births, deaths, and property ownership captured in public records, or birth dates and address information captured in customer records (Connelly, Playford, Gayle, & Dibben, 2016). The next most entrenched/longstanding form of TM data is consumer data that can be obtained from a wide range of transactional records, such as purchase details, loyalty cards, and product/service usage (Birkin, 2019). In recent developments, some TM databases are integrating digital data that track individuals' online browsing patterns and access. Relatedly, another form of online data is derived from social network platforms, and may include information ranging from contact networks to taste preferences regarding movies, news content, music, etc.

In addition to these passive data sources, TM data may also include active data sources from surveys that are typically conducted by consumer research firms, but which can also come from individuals' responses to online quizzes/games/ questionnaires. For clarity, it is noted that to qualify as active data, the individual must choose to relay the information being obtained, while with passive data, the individual may not even be aware that information is being collected. TM databases often comprise information from both active and passive data sources, a characteristic differentiating them from traditional third-

party and/or big data, which are typically entirely derived from passive data sources. The TM variables that are derived from active data sources like surveys typically include individuals' preferences and opinions toward specific products and/or services (e.g., the importance of post-purchase customer service in selecting a specific type of service), but can also include more general preferences. Examples of the types of variables present in TM databases can be found in Section 2.5.1 and Table B1 of Appendix B.1.

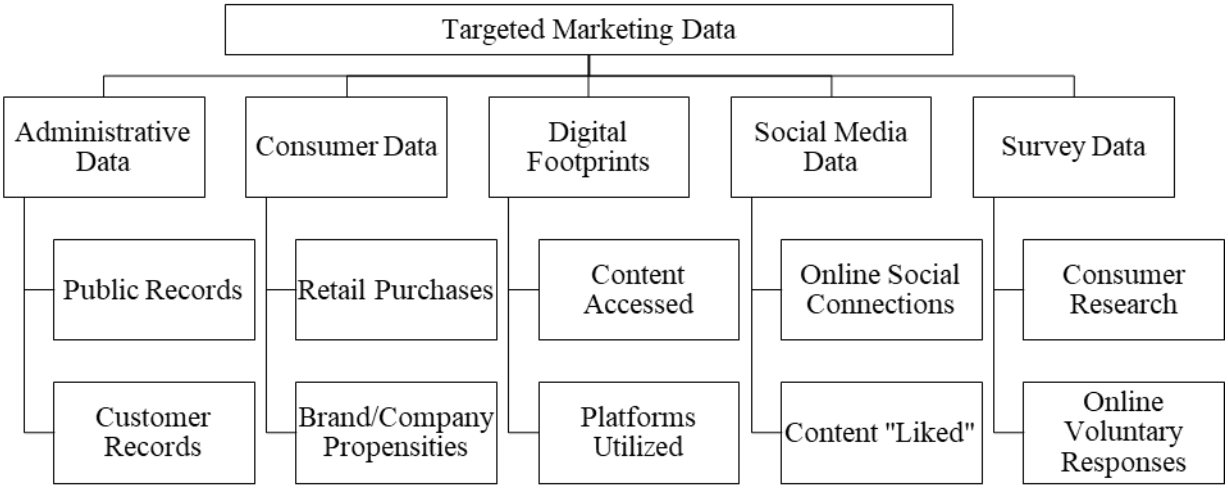

Figure 2.1. Examples of TM data sources

2.3.2 Benefits of targeted marketing data

The most significant benefits of TM data within a transportation context are the volume and disaggregate nature of the data. TM datasets are extremely large because they are available for almost all individuals/households in the population, allowing for the possibility of using TM data to enrich other data sources at a disaggregate level for most individuals in a typical transportation study. This contrasts with most publicly available data (e.g., Census and American Community Survey), which are commonly used for transportation data validation, but which report only aggregate-level cross tabulations (e.g.,

block groups, census tracts) or are available only for a small fraction of the population (e.g., the Public Use Microdata Sample). The resulting potential magnitude of TM data may also facilitate the use of efficient artificial intelligence approaches for researchers and practitioners interested in using these methods.

Further, in contrast to traditional large-scale household travel surveys which tend to occur once every 10 years or so, TM data is dynamic, meaning that the values of many TM variables are updated on a monthly, quarterly, or yearly basis. In addition, TM data comprise a range of diverse variables, many of which are not available through traditional or novel data sources currently used in transportation. Thus, the primary overall benefits of TM data lie in the overall magnitude/size of the data, the diversity/richness of TM variables, and the rapidity of TM data generation and renewal (Erevelles et al., 2016; Sivarajah, Kamal, Irani, & Weerakkody, 2017). These three attributes are respectively known as volume, variety, and velocity, and also happen to be considered the original three defining attributes of big data (Laney, 2001). “Value” and “veracity” were added later, with these five attributes collectively being known as the “5 Vs” commonly used to characterize and evaluate big data (although, note that one could find lists ranging from seven to forty-two Vs that are used in various contexts to further describe big data; Sivarajah et al., 2017). However, while TM databases have increased volume, variety, and velocity relative to transportation survey data, they are generally smaller in size and are generated more slowly than traditional big data, which tend to be purely passive data such as second-by-second GPS location traces. Nonetheless, relative to traditional transportation survey data, TM data can be considered to meet the loosely defined and broadly applied definition of big data (Macfarlane, 2014).Regarding utilitarian benefits, TM data are inexpensive and easily

accessible relative to traditional survey data collection. For this work, TM data cost approximately $1.50 per person, while a statewide transportation survey (i.e., GDOT survey) that obtained rich attitudinal and behavioral variables cost an estimated $19.85 per person. Further, the purchased TM dataset contained 5583 variables, while the transportation survey dataset contained 200 – 400 unique variables (using varied coding techniques). The overall cost of the transportation survey was ~$65,000, while the overall cost of purchasing the TM dataset for more than three times as many respondents (~10,000 cases) as were contained in the survey final sample (~3,000 cases) was ~$15,000, not counting graduate student/faculty time costs for either. At a household level, Kressner and Garrow (2014) reported that in their estimates, the cost of obtaining a completed travel survey for one household in Atlanta is around $200, relative to five cents for obtaining a set of TM variables for that household (the study did not detail *how many* TM variables were obtained). Finally, a significant benefit for transportation professionals is that TM data have widespread availability, meaning that any entity, from academic researchers to governmental agencies, could purchase TM data from marketing firms (after agreeing to legally mandated privacy restrictions). This accessibility means that if TM data are shown to improve modeling/forecasting, transportation agencies can feasibly acquire TM data and integrate them into their operations; however, as will be discussed in more detail next, with increased privacy restrictions, this availability may be moderated in the future.

2.3.3 Challenges of targeted marketing data

Before TM data can achieve widespread utilization in transportation, it is important to assess the value (i.e., worth/usefulness) and veracity (i.e., accuracy) of these data within the context of intended applications (Lavalle, Lesser, Shockley, Hopkins, & Kruschwitz,

2011; Lovelace, Birkin, Cross, & Clarke, 2016; Lukoianova & Rubin, 2014; Sivarajah et al., 2017). While a handful of studies have shown the value of TM data in transportation (detailed in Section 3), to date, only three studies are known that have sought to examine the veracity of TM data from a transport perspective (Kressner & Garrow, 2014; Kressner, Carragher, & Watkins, 2014; Lovelace et al., 2016). This may be partially due to challenges associated with integrating TM variables with traditional travel datasets, namely that names and addresses are needed to obtain TM data for individual-level validations; however, this does not restrict aggregate level validations, which are similarly rare.

Further underscoring the importance of evaluating the value and veracity of the data is the fact that, as with all data sources, TM data have inherent biases that may disproportionately affect underrepresented/ vulnerable populations. To begin the process of mitigating these challenges, this chapter provides a guide to integrating TM data with existing transportation datasets, and further presents both an individual/ household-level pairwise validation (Section 2.6.1) and an examination of TM data biases and representativeness (Sections 2.6.2 and2.6.3). Further, Section 2.7 provides a brief discussion of methods for ameliorating dataset biases that may be useful in the specific context of the TM data being examined in this study.

A second set of challenges in working with TM data lies in the development of the variables. TM providers often use proprietary algorithms to develop, impute, and/or model many variables, not only making it difficult to evaluate the robustness of TM variables, but further clouding the interpretation of these variables if they are to be used in transport models. It is emphasized here that this constitutes a significant disadvantage of third-party data sources like TM data relative to first or second-party data that are often more

transparent regarding variable development procedures. In addition, modeled TM variables may be relatively unstable as the algorithms may be tweaked over time, thus precluding consistent definitions of the variables. Furthermore, variables themselves may become obsolete as the data sources used to inform the TM databases ebb and flow, in part in response to the commercial demand for the associated information. Moreover, variables are both measured on different time frames and updated on a schedule that differs across variables and which may not be transparent to the user. For example, a variable indicating whether the individual has purchased a car within the past 12 months may have been last updated 11 months ago (and therefore be almost a year out of date), while a variable indicating whether the individual has had food delivered to the house within the past month may have been last updated six months ago.

Nonetheless, such issues are present in most external data sources, as variable definitions and included variables change even across national data sources such as the U.S. Decennial Census and National Household Travel Survey. Moreover, these challenges do not detract from the richness of the information that TM data have to offer, and in reality, there are numerous consistent TM variables that users can rely on while avoiding variables that may be unclear or unstable. Furthermore, as with most big data, when methodologies like machine learning are used, the stability and interpretation of variables are arguably less important than their contribution to an overall improvement in forecasting that facilitates more accurate decision making. In Section 2.4, it is shown that post-model development, TM can be used to develop policy scenarios, thereby compensating for the reduced interpretability of some variables in model development.

From another perspective, the quantity and richness of TM variables provide an added challenge. Specifically, since TM data come from a large array of sources, there may be reduced consistency in data scales and definitions across variables, as was experienced here. As such, users may have to spend additional time processing the received data, and in some cases, building their own data dictionaries. Thus, as acquired data become increasingly voluminous and diverse, the potential to obtain value is moderated by the available physical, human, and organizational capital (Sivarajah et al., 2017). It is also worth noting that since TM data is collected and aggregated for marketing purposes, the resultant databases do not contain the same breadth of general and transport-related preferences and opinions that can be obtained using transportation survey data. The challenges discussed here are likely some of the major reasons slowing the use of TM in transportation, and it is hoped that this work, in combination with additional efforts from other TM data users, will serve to introduce the requisite outlook and approaches needed to overcome these challenges.

The final group of challenges for using TM data are evolving privacy regulations and concerns that are increasingly salient to researchers, regulatory agencies, and the public. The European Union (EU) General Data Protection Regulation (GDPR), introduced in 2018, represents the strictest data protection law in the world to date. Even though the U.S. as a whole is currently far from this level of regulation, some states have expressed interest in emulating the EU, such as California, which instituted the California Consumer Privacy Act (CCPA) at the start of 2020. While the specifics of these laws are complex, the most relevant detail in the context of this chapter is that both laws aim to provide consumers with the ability to opt out of the collection and sharing of their personal

information, and/or to edit the consumer records that are available to them. In practice, this means that TM data providers will still retain existing databases identical to those described in Section 2.3; however as mentioned, consumers can request corrections or deletions made to their records that are present in those databases (Acxiom, 2020). At this point it remains to be examined how this new provision will affect TM databases for European and Californian consumers in the future, a point that of course rests on how many consumers take advantage of the policies to remove/edit their records in the database. Future research should seek to explore the changes that have occurred in the databases as a result of new privacy laws.

Thus, overall, from a data availability standpoint, evolving privacy regulations may threaten the stability and reliability of TM data for long-term transportation applications, particularly those that require TM records to be matched at a disaggregate level. Despite these complications, third-party data such as TM data are expected to continue to be critical supplementary data sources for a wide array of fields, and as such, this document aims to provide a stimulus for transportation professionals to explore compliant and ethical approaches to using these diverse data sources to improve transportation modeling and forecasting efforts. One potential solution may lie in the use of data agencies that can serve as intermediaries between data providers and researchers, thus ensuring that the data provided to individual research teams has been appropriately processed to prevent any potential privacy incursions (examples of agencies that could/ already serve this purpose are the United Kingdom Administrative Data Research Network, the Consumer Data Research Centre, and the University of Washington Transportation Data Collaborative). Regardless of how the data are acquired, it is recommended that analysts meet with

appropriate institutional research ethics personnel prior to beginning any project that uses third-party data sources, and once the data have been acquired, to work toward timely de-identification of the datasets being used.

2.4 Using targeted marketing data

TM data can be used in each stage of travel demand modeling and forecasting efforts, beginning with survey design and sampling, extending to model prediction and accuracy, and even having implications for result interpretation and application (see typology in Figure 2.2). In this section, it is shown that the outcomes and research directions associated with TM data in transportation are non-trivial, and have the potential to significantly improve transport planning in the future. Where applicable, examples of known TM applications in the transportation literature are cited.

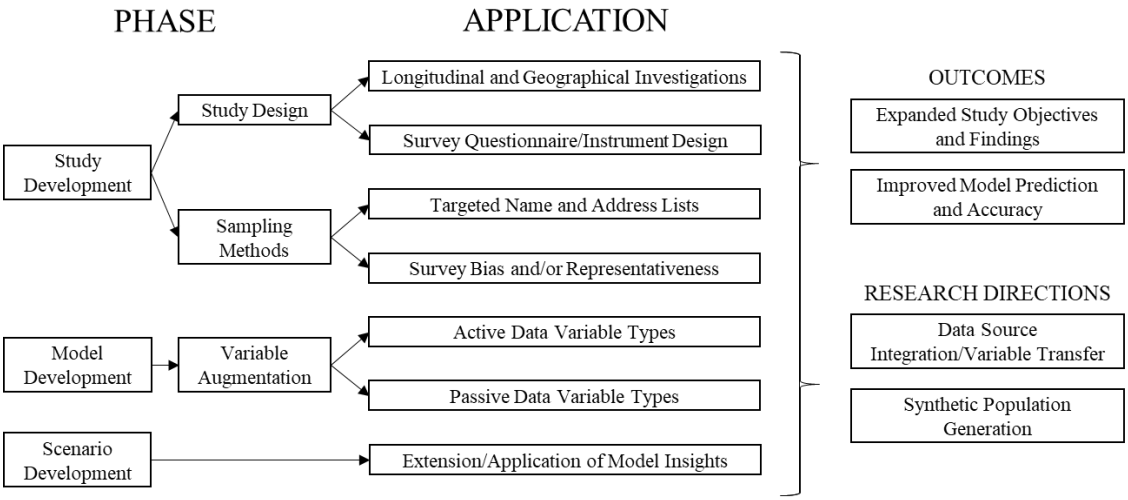

Figure 2.2. A typology of TM applications in transportation

2.4.1 Transport applications and outcomes

TM data have significant implications for shaping transport study development, from both study design and sampling perspectives. The dynamic nature of TM data is one of its most significant benefits with regard to study design, as this allows for ease of data collection at multiple timepoints. Documented changes in individual/household characteristics over time may allow for improved understanding and forecasting of how these changes influence travel behaviors. Macfarlane et al. illustrated the benefit of the longitudinal nature of TM data by using address histories from TM data to examine how prior places of residence could influence vehicle ownership, a study objective that would not be possible with traditional cross-sectional survey data (Macfarlane, Garrow, and Mokhtarian, 2015). Birkin (2019) later similarly suggested that consumer data (in this case, from online real estate agents) is unique in providing the level of spatial detail (i.e., origins and destinations) and rapid updating necessary for the study of geodemographic mobility, a key transport geography study objective that previously required the use of longitudinal data. In the same way that TM data are present across time, they are also available across regions, facilitating geographic and land use comparisons for transportation attitudes and behaviors, and further, providing the ability to validate/ segment models along those lines.

From a survey instrument design perspective, the presence of TM data for respondents being sampled could aid in the reduction of the number of questions necessary on travel behavior surveys, thus resulting in shorter surveys and thereby, potentially improved survey response rates. Alternatively, if some variables are able to be reliably sourced from TM data, then the corresponding survey questions may be replaced with other

questions, yielding a richer set of variables for use in model building, and thus potentially improving model predictions.

TM variables can also aid sampling efforts for traditional transportation data collection. TM databases are already widely used to obtain names and addresses for use in travel behavior survey sampling (e.g., Handy, Cao, & Mokhtarian, 2005; Kressner et al. 2014). Building on that, the SED characteristics present in TM data (e.g., gender, income) can allow analysts to sample socioeconomic and demographic (SED) groups of interest in greater proportions relative to other groups. For example, it is known that individuals in certain SED groups have lower or higher response rates relative to other groups, and respondent information based on the TM variables could aid in over-/under-sampling as appropriate. TM data could also be used to examine survey biases and representativeness by providing an additional source of information that could be compared with traditional survey data, although it is noted that traditional survey data and TM data will each have their own inherent biases, a point further examined in 2.6.

TM data have perhaps the greatest potential to benefit transportation model development through the augmentation of transportation datasets with variables that are not possible to obtain through traditional transportation surveys (i.e., passive data), as well as variables that are obtained through active data collection methods. As a result, given appropriate prior hypotheses, TM data can facilitate the testing of a larger range of variables in predictive models, leading to enhanced conceptual understanding of travel-related attributes, as well as potential improvements to model performance. Some transport studies have already shown that unique TM variables can improve model accuracy; for example, Kressner showed that using TM lifestyle segments improves prediction for air

passenger trip models and residential location choice models (Kressner and Garrow, 2012; Kressner, 2014). In addition to serving as explanatory variables, in some cases active and passive TM variables may also be of interest as dependent variables (i.e., variables to be modeled in their own right) within larger travel demand models/systems. Furthermore, the distributions or frequencies of TM variables of interest may also be used to aid in model development, serving to provide marginal distributions or probabilities that could potentially aid in model calibration (although of course it will be critical to first ensure representativeness of the data being used – see Section 2.6 for more on TM data representativeness).

On the other hand, Binder, Macfarlane, Garrow, and Bierlaire (2014) showed that TM variables typically obtained through survey data collection, such as ethnicity, income, gender, and age, are able to support residential location choice models without depending on HHTSs. This significant finding could allow researchers not only to shorten their surveys, but also to remove more sensitive questions (e.g., income) from survey instruments, both actions which could allay some of the contributors to declining survey response rates. In a similar example, Macfarlane, Garrow, and Moreno-Cruz (2015) used SED traits and home prices derived from TM data to model willingness to pay for proximity to public transit. In addition to these examples in the literature, many regional transportation planning agencies also currently obtain employment statistics (for use in their regional models) from business list data acquired through TM firms. Overall, the outcomes possible from augmenting traditionally available travel datasets with TM data offer significant implications for the field, and it is for this reason that Section 2.5 of this chapter provides a generalized framework that can aid in pursuing this application.

Lastly, TM data are ripe for use in the development and testing of policy scenarios, applications that can expand the insights gleaned from analyses, while potentially clarifying decision-making based on transport study findings. Specifically, TM data can facilitate the post-hoc application of models and/or proposed policies to various segments of the population, allowing for an understanding of how proposed scenarios may affect individuals, demographic groups, overall transport choices, and infrastructure operations. Furthermore, TM data can be purchased for this purpose even after the completion of a study, thus making TM integration at this stage more accessible. In one example from the literature, Binder et al. (2014) used data derived from TM records to examine the effects of three proposed emissions policy scenarios on various SED groups, finding that the suggested and commonly used strategies for reducing the cost of indiscriminate emission testing are inequitable and/or ineffective, and suggesting that other transportation policy tools may be needed to address the issue.

2.4.2 Transport research directions

The preceding section highlighted the potential for TM data to expand transportation study objectives and improve model predictions. Beyond these outcomes, there are many transportation research directions that could benefit from the use of TM data. Two such examples involve the use of methodological tools like machine learning and discrete event simulation to aid in: (1) the integration of multiple data sources through variable transfer; and (2) the generation of synthetic populations based on disaggregate TM data.

The first initiative is actually developed and presented in Chapter 3 of this document, using the integrated dataset described in Section 2.5. In this effort, a range of algorithms are trained using an integrated dataset that combines statewide and nationwide transportation surveys with TM data appended at an individual/ household level. High performing algorithms that are able to predict selected variables (e.g., attitudes) may facilitate the transfer of variables that are unique to one data source into a recipient data source. This approach paves the way for data source linkage, with TM data operating as the "glue" (i.e., "common" variables/features) that links disparate sources together, and facilitates variable transfer. This approach may enable the development of richer, more up-to-date datasets that can improve travel demand modeling efforts.

The second group of initiatives entails the use of disaggregate TM data to generate synthetic populations that can yield insights into how individuals in a region travel (Beckman, Baggerly, & McKay, 1996; Birkin, Morris, Birkin, & Lovelace, 2017; Kressner, Macfarlane, Huntsinger, & Donnelly, 2016; Kressner, 2017). The use of disaggregate TM data to provide a nearly-complete enumeration of household and individual-level SED traits may represent an improvement over the 1% or 5% anonymized sample offered by the American Community Survey Public Use Microdata Sample (ACS PUMS), which is currently the primary source of SED inputs for population generation in transportation. Kressner has implemented this idea at a large scale, using TM data to provide disaggregate SED data that is then fused with mobile phone location data to create synthetic travel diary records (2017). This concept has been successfully validated for several cities in the U.S. (Kressner et al., 2016). Along similar lines, researchers in Europe simulated demographics that would match Census data for a city, and then matched travel-related consumer data to

these simulated individuals on the basis of age, gender, family status, and social group (Birkin, Morris, Birkin, & Lovelace, 2017).

Thus, while the first research initiative detailed here uses TM data to allow variable transfer across data sources, the second approach uses it to synthesize populations, and study how these synthetic populations travel. Both approaches highlight the importance of integrating passive and active data sources to build and validate disaggregate/aggregate travel demand modeling systems, a tactic that can help take travel demand modeling into the next generation by reducing the reliance on traditional data sources. While the ultimate effectiveness of these approaches, and possible symbiosis of methods, remains to be seen, it is believed that there is substantial potential not only in these methods, but also in future approaches that can use TM data to make similarly ambitious attempts to move the field forward.

2.5 Integrating targeted marketing data

As discussed in Section 2.3.3, to examine the value and veracity of TM data for use in transport applications, TM data must first be integrated with transportation survey datasets. However, the integration of TM data with other data sources can pose technical and methodological challenges. As a result, in the following subsections, an overview of the TM data used in this study is provided (Section 2.5.1), followed by a discussion of the process used to integrate TM data with transport survey datasets (Section 2.5.2, with additional details in Appendix B.2).

2.5.1 Overview of targeted marketing data used in study

The TM data purchased for use in this study was obtained from a large U.S.-based TM data provider that is an industry leader in data quality, and which is used by many business entities for their marketing needs. Selection of the provider used in this study hinged on the firm's ability to provide a rich array of variables for the smaller sample size (~10,000 cases) and nontraditional (exploratory, research-based) data needs of this project. In addition to the TM firm's natively collected/derived variables, their database also houses supplementary variables purchased from well-known firms such as Claritas, SEMcasting, etc. At the time of acquisition for this study, the firm's database contained ~5500 variables ('p' is used to represent number of variables throughout this document), all of which were purchased for this study.

Of the total variables available, approximately 1500 represent a general variable set from which most marketers (i.e., typical clients for TM firms) select when purchasing data augmentation services. The additional ~4000 variables are termed audience propensity variables, and are developed on contract to be sold to certain corporations, and thus might be updated/changed on a monthly basis. The general variables have no name release restrictions, meaning that the full names can be shared publicly, while the audience propensity variables required a legally binding non-disclosure agreement barring disclosure even of these variables' names. Further, to obtain the full set of all variables, we provided an official statement of use followed by the completion of additional legal paperwork on the terms of use for these variables. Certain variable subsets (such as sensitive financial variables) required the TM provider to obtain specific approval from the firms that generated those variables before they could be included in the overall purchase

for this study. Thus, as can be seen, the process of obtaining a *large* TM variable set is a non-trivial undertaking that can require months of discussion prior to final approval and variable transmission.

The acquired TM dataset comprises continuous, ordinal, and nominal (dichotomous and polytomous) variables. In Figure 2.3 and Table B1 of Appendix B.1, the initial received variables (after removing variables that were completely missing, as well as meta-data variables like precision levels) are classified into the following topical areas: sociodemographic, land use, attitudes, lifestyle, financial, technology, and transportation. Figure 2.3 summarizes the overall variable distribution, and Table B1 further summarizes the variable classification distribution across the TM variables. Given the traditional TM sources of credit card and shopping records, it is intuitive that 61% of the received TM variables are consumer-related variables such as purchase behavior, while 18% are financial variables related to investment, income, and insurance, among others. Examples of transportation variables obtained include business and vacation travel behaviors, vehicle ownership (i.e., brands/vehicle type), vehicle payment type, and brand propensities regarding rental car companies and airlines.

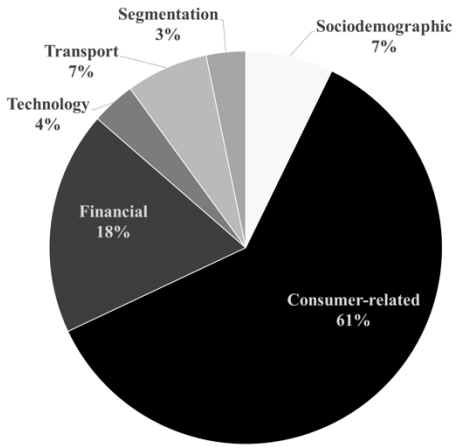

Figure 2.3. Overview of variable types in TM dataset (p = 5684)

2.5.2 Targeted marketing data integration framework

In Figure 2.4, the process of acquiring, processing, and integrating TM data with existing transportation survey datasets is summarized (bolded elements depict the process used in this study). In this section, only a brief overview is provided, but interested readers are directed to Appendix B.2, where an in-depth guide with step-by-step detail on the integration process used for the datasets in this document has been provided.

There are four primary services of interest offered by TM data providers, and transportation professionals may be interested in any of these services for varying applications (see Section 3). In this section, *data enrichment* only is discussed, as it is the service used to append a range of TM variables to existing records (see Section B.2.1 of Appendix B.2 for a discussion of all data services). To use this service, analysts should first determine the quantity and types of TM variables intended to be appended to each record. For a small number of variables (i.e., 50 – 100), TM providers often have online portals that can be used to quickly and easily append variables. As the number of variables

and/or respondents grows, data enrichment must proceed through in-house services that require additional legal paperwork and time.

TM providers typically require names *and* addresses for all cases that are being submitted for TM data enrichment. Submitted lists are matched against names and addresses on file in the TM provider's database, and if the exact first and last name cannot be matched, variable matches degenerate into less precise matches (e.g., address and last name, address only, zip+4 code, zip code – with each of these successively identifying a larger, less precise area where for example, zip+4 code may refer to a specific part of a street or a building while zip code may refer to a general area and/or associated mail delivery office). Since transportation practitioners may have varying amounts of name/address information available for their survey datasets, in Section B.2.2 it is shown how the four survey data subsets in this study were approached (Figure 1.2), as each had differing amounts/types of name/address information available.

Following data acquisition, the resulting TM dataset typically requires substantial cleaning, recoding, and processing before integration with survey datasets. The most critical step entails the individual-level comparison of the TM record for each case to the available survey data. Analysts must first select the variables that will be compared between the TM and survey data, and subsequently should establish the associated tolerance/confidence level for retaining the compared cases given the selected variables. For the dataset in this chapter, the variables selected for verification are gender, age, and education level, in order of importance. After processing and retaining cases that are believed to represent the same individual across datasets, TM variables must be recoded (e.g., variable values/levels may need to be made consistent across data sources), cleaned

(variables with high levels of missingness or near zero variance may need to be removed or otherwise addressed), and imputed as necessary.

As before noted, please reference Appendix B.2 (Section B.2) for expanded guidance on selecting a TM provider and service, successfully acquiring TM data, and cleaning and processing the obtained dataset.

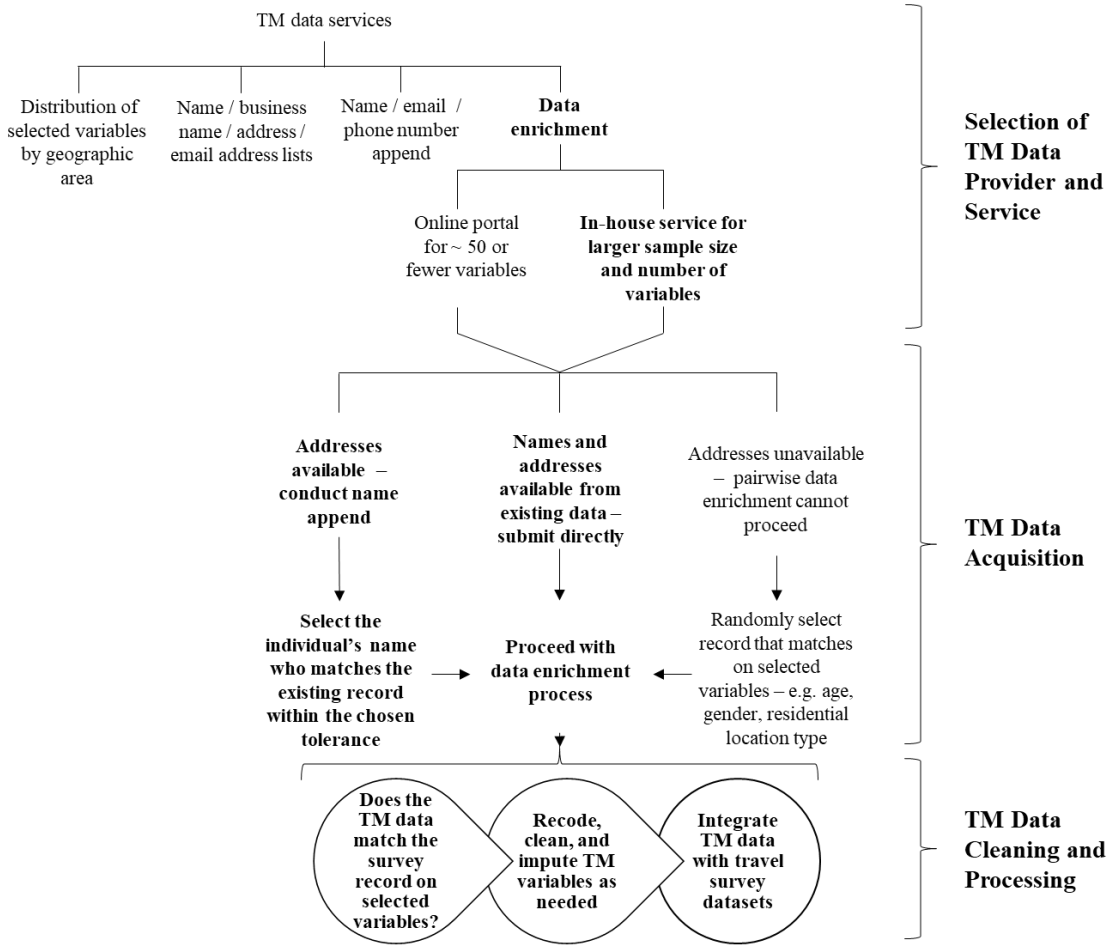

Figure 2.4. Simplified overview of TM data integration process

2.6 Validating targeted marketing data

At this point, the TM variables purchased for this study (summarized in Table B1) have now been successfully integrated with two transportation survey datasets (Section 1.3) using the framework and approach outlined in Section 2.5.2 and further detailed in Appendix B.2. Following this data integration step, comparable variable categories are developed for fundamental SED variables present across the TM data and survey datasets (see Table B2). The discussion turns now to validating these key TM variables relative to the GDOT survey and NHTS datasets discussed previously.

Statewide/regional surveys (which include research-oriented surveys like the GDOT survey, as well as regional household travel surveys), in tandem with nationwide data sources like the Census, American Community Survey (ACS), and NHTS, represent the core sources of data used in transportation planning and forecasting. Thus, examining TM data relative to these transportation surveys, and further, being able to compare the NHTS and GDOT surveys relative to each other, represent unique contributions of this chapter.

2.6.1 Investigating differences between TM and travel survey variables

The first step in assessing the quality of TM data lies in verifying the "accuracy" of its values for critical variables, such as SED variables, in the TM database. An ideal approach would entail the validation of TM variables with values from official records or reports (or alternatively, in-person verification). Given the absence of reliable SED data from publicly available disaggregate personal records, as well as the focus of this chapter on examining TM data within a transport context, select TM variables are validated based

on corresponding variables obtained/derived from the GDOT survey and the NHTS. Comparing TM data to federal and statewide transportation survey data can help transportation practitioners to better understand whether it is possible to replace, augment, and/or model specific travel behavior survey data with TM data, and further, can provide guidance for addressing identified discrepancies.

To date, the only TM validation studies for SED variables of which the authors are aware include an aggregate level validation for TM data at the block group level (Kressner and Garrow, 2014), as well as a small-scale household-level validation between TM and travel survey data (Kressner et al., 2014). As findings from these prior studies will be compared to results from the analysis in this chapter, it is pertinent to note that the household-level validation in Kressner at al. (2014) used survey data from hard-to-reach populations, thus indicating a bias in the survey data toward individuals living below the poverty line.

Accordingly, the data validation presented in this section extends the preceding investigation by: (1) expanding the household-level validation to significantly larger (from $N = 116$ to $N \approx 5000$) and more representative samples; (2) allowing for the simultaneous pairwise comparison of TM data with two different types of transportation surveys; and (3) illustrating the effect of TM data processing on variable match rates. To facilitate comparison of the validation process with Kressner et al. (2014), a match (on a given variable) between the same case in two different datasetsis defined as being accurate if the case has the same value (within a tolerance band, if applicable) for that variable in both samples, and inaccurate otherwise. For example, if a given individual is in the 18-24 age category in the NHTS survey, but in the 25-34 age category for the TM data, then that case

is considered an inaccurate match on age. The shares (or, if expressed as percentages, rates) are calculated only on comparable cases, as follows:

$$Accurate\ match\ share\ = \frac{Number\ of\ comparable\ cases\ with\ same\ variable\ value}{Number\ of\ comparable\ cases}$$, and

$$Inaccurate\ match\ share = \frac{Number\ of\ comparable\ cases\ with\ different\ variable\ value}{Number\ of\ comparable\ cases}$$,

where "comparable cases" refers to cases that were able to be assigned a value that was able to be developed across all data sources (Table A3). Noncomparable cases include those with missing, not applicable, not able to be classified, other, "I don't know", and "prefer not to answer" responses to the variable in question in one or both datasets being compared. Using these definitions, Table 2.1 and Figure 2.5 summarize match rates across the entire TM and survey datasets used, while Table B3 and Figure B1 in Appendix B.1 summarize these rates for the *same* respondents across all three datasets (i.e., for the overlapped sample). Table B3 also includes GDOT/NHTS variable match rates to allow for insight into differences between the surveys. Prior to comparing the variables selected for validation, it was necessary to recode several variables into directly comparable categories; Table B2 in Appendix B.1 summarizes this process, and details final variable values used. For consistency, values for the NHTS, GDOT, or TM variables are not imputed by the authors; however, some of the TM variables were imputed/infilled prior to our receipt of those variables. The TM variables that were specified as imputed in the TM database include household income and household size variables, which had missing values filled in with zip code and/or zip+4 code data, and the marital status variable. which was filled in with undisclosed imputations. We note that other TM variables may have also

been imputed in some way, but those listed here are the ones that were transparently listed as having been imputed in the TM database documentation.

As illustrated in Figure 2.5 and Table 2.1, the match rates when comparing the TM and survey data are generally consistent for both the NHTS and the GDOT survey, with the highest accuracy rates occuring for gender, age, tenure, race, marital status, and dwelling type, and the lowest accuracy rates occuring for occupation, income, education, and household size. It is seen that gender has the overall highest percentage of accurate matches for both NHTS and GDOT data (90.9% and 95.6%, respectively), followed by age with match rates of around 89-91%. It's posited that gender and age may have the highest match rates between TM and survey data due to the ease of obtaining these variables from publicly available records (e.g., birth records), although gender identification is also believed to be derived based on typical male and female names in the Caucasian population. This latter proposition is based on the observation that foreign names (e.g., names of Asian or Native American origin) are often listed as unidentifiable with regard to gender. Race had accuracy rates of ~85% for both surveys, representing the fourth highest match rate among SED variables examined.

Housing tenure was comparable between NHTS and TM data only, and had the third highest accuracy rate of 87.31%, while marital status and dwelling type were only comparable between GDOT and TM data, and had the next highest accuracy rates of 72.2% and 63.05%, respectively. Occupation had lower accuracy rates of ~59% between GDOT and TM data and ~55% between NHTS and TM data; however this is likely because ~75% of the cases could not be compared. While dwelling type and occupation were not studied in prior literature, it is noted that for gender, tenure, and marital status, the findings shown

here are consistent with those of Kressner et al. (2014). Regarding age and race, significantly higher accuracy rates are identified here relative to the prior work, potentially suggesting either a bias in TM data in reporting these variables for under-represented populations, or that the TM database used in this study had more accurate data on age and race.

Rounding out the rest of the variables, income, education, and household size all had accuracy rates below 55%, findings also shown by Kressner et al. (2014). This indicates the robustness of the finding that these individual-level variables have low accuracy rates (i.e., almost 50% or lower) in TM data, since they continue to do so five years after an initial validation study. Such performance may be attributable to the relative transience of these variables; for example, income, education, and occupation can all change several times over an individual's lifetime (we note as well that these variables do not change *consistently* over time relative to a transient variable like age). Similarly, household size is a constantly in-flux variable, as individuals marry/divorce/die and give birth to children, and as children move out of/into the household. When a tolerance of ±1 is allowed when calculating the household size accuracy rates, it is seen that the match rates more than double (from ~30% to ~70%, for both categorical and continuous versions of the variable), supporting the conjecture that for dynamic variables, TM may take several months to years to receive updated information, which at least partially accounts for the low accuracy rates observed. Thus, it is worth noting that for a low performing variable like household size, TM *can* provide more accurate estimates *within* a certain tolerance.

As discussed before (Section 2.5), gender, age, and education were used to process the TM data to retain records that were believed to correspond to the correct indvidual in

the survey data sources. As Table 2.1 and Table B3 (Appendix B.1) show, even after data processing (i.e., the sample is filtered to include only individuals who are considered to be definite matches between the TM and survey data), all of the variables with the exception of age and gender saw only small improvements in accuracy, suggesting that the accuracy rates observed for race, marital status, dwelling type, occupation, income, education, and household size are largely representative of the rates that could be typically expected for such variables in TM databases.

In the next sections, (2.6.2 and 2.6.3) additional validation approaches are explored. These include first examining distributional differences in the accurate and inaccurate matches, followed by a modeling effort that examines the factors influencing individuals' propensities to be matched correctly in the TM database.

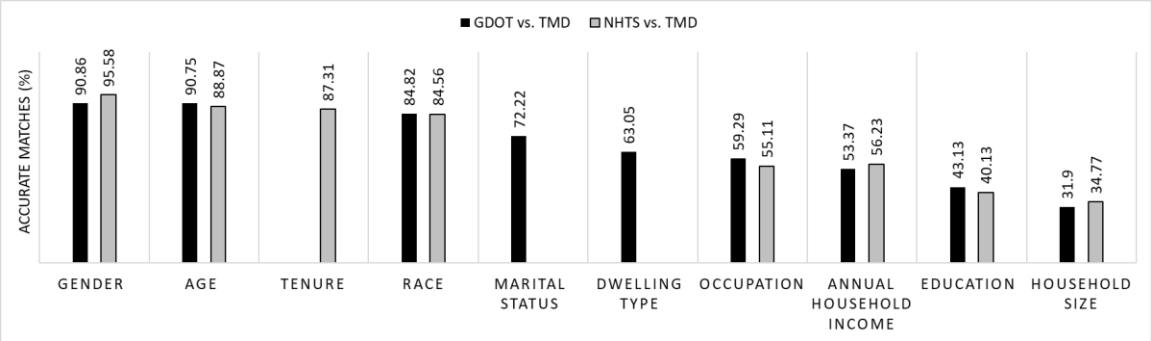

Figure 2.5. Variable accuracy rates across TM and survey datasets *before processing*

Table 2.1. Variable accuracy rates across TM and survey datasets before and after processing

Variable	Match	Before Data Processing				After Data Processing			
		TM vs. GDOT N = 3288 [a]		TM vs. NHTS N = 5148 [a, c]		TM vs. GDOT N = 2699 [b]		TM vs. NHTS N = 4027 [b, c]	
		N	%	N	%	N	%	N	%
Gender [f]	Accurate matches [d]	2864	90.86	4455	95.58	2686	100.00	4019	100.00
	Inaccurate matches [d]	288	9.14	206	4.42	0	0.00	0	0.00
	Not comparable [e]	136	–	487	–	13	–	8	–

Table 2.1 cont'd

Age [f]	Accurate matches	2806	90.75	4023	88.87	2610	99.35	3710	95.10
	Inaccurate matches	286	9.25	504	11.13	17	0.65	191	4.90
	Not comparable	196	–	621	–	72	–	126	–
Tenure [g]	Accurate matches	–	–	4168	87.31	–	–	3519	88.06
	Inaccurate matches	–	–	606	12.69	–	–	477	11.94
	Not comparable	–	–	374	–	–	–	31	–
Race	Accurate matches	2441	84.82	3708	84.56	2020	85.59	2967	85.53
	Inaccurate matches	437	15.18	677	15.44	340	14.41	502	14.47
	Not comparable	410	–	763	–	339	–	558	–
Marital status [g]	Accurate matches	2111	72.22	–	–	1807	73.85	–	–
	Inaccurate matches	812	27.78	–	–	640	26.15	–	–
	Not comparable	365	–	–	–	252	–	–	–
Dwelling type [g]	Accurate matches	1635	63.05	–	–	1348	61.89	–	–
	Inaccurate matches	958	36.95	–	–	830	38.11	–	–
	Not comparable	695	–	–	–	521	–	–	–
Occupation	Accurate matches	498	59.29	701	55.11	455	61.57	641	56.08
	Inaccurate matches	342	40.71	571	44.89	284	38.43	502	43.92
	Not comparable	2448	–	3876	–	1960	–	2884	–
Annual household income	Accurate matches	1686	53.37	2852	56.23	1418	54.62	2215	55.82
	Inaccurate matches	1473	46.63	2220	43.77	1178	45.38	1753	44.18
	Not comparable	129	–	76	–	103	–	59	–
Education [f]	Accurate matches	1167	43.13	1560	40.13	1092	47.44	1456	43.41
	Inaccurate matches	1539	56.87	2327	59.87	1210	52.56	1898	56.59
	Not comparable	582	–	1261	–	397	–	673	–
Household size [h]	Accurate matches	1049	31.90	1790	34.77	879	32.59	1396	34.67
	Inaccurate matches	2235	68.10	3358	65.23	1818	67.41	2631	65.33
	Not comparable	4	–	0	–	2	–	0	–

[a] An overlap sample of 1495 respondents exists in the NHTS and GDOT survey datasets before processing.
[b] An overlap sample of 1245 respondents exists in the NHTS and GDOT survey datasets after processing.
[c] Respondents who *did not want to be contacted again* are removed from the NHTS samples, as this subset had TM pre-processing prior to data enrichment. See Section 2.5 and Appendix B.2 for more information.
[d] Match percentages exclude "Not comparable" segments and should be interpreted as the percentage of respondents who could be compared with an equivalent category between data sources that are accurately matched (or inaccurately matched). Table B2 in Appendix B.1 summarizes the variable values that are compared to each other.
[e] The "Not comparable" value includes respondents in "Other/Could not be classified/Not applicable/Prefer not to answer/Missing" categories. These categories were not separated, because they are often confounded across sources. For example, in the TM data sources, "Missing" and "Not applicable" were not distinguishable from each other, although they were distinguishable for some of the questions in the survey data sources.
[f] Gender, age (tolerance +/- 4 years), and education (tolerance: +/- 2 levels) are used in post-processing to ensure that the TM records obtained are appended to the correct individuals. As such, the accuracy for these numbers in the post-processed sample are higher than would be typically expected (or unrealistically perfect, as in the case of gender). Note that even when instituting these matching criteria, we were able to retain 82.09% of the GDOT respondents and 78.22% of the NHTS respondents (i.e., we are relatively confident of having the correct TM records for ~80% of survey respondents). There remain "Not comparable" cases for gender, age, and education in the post-processing sample because we retained cases for which gender/age/education are missing in either the TM or survey datasets, as these could not be definitively ruled out based on inaccurate matches.

Table 2.1 cont'd

[g] NHTS did not obtain marital status and home dwelling type of survey respondents, and thus these variables could not be compared between TM and NHTS data. Similarly, GDOT did not obtain tenure, and thus this variable could not be compared between TM and GDOT survey data.
[h] When a tolerance of +/-1 was instituted for the household size variable, the percentage of accurate matches increases substantially, to: 71.92%, 72.18%, 72.38%, 72.29%, in respective order of the four percentages listed in the table.

2.6.2 Examining patterns in response differences between TM and survey data

This section examines whether the distributions of variable accuracy and inaccuracy are associated (i.e., correlated) with the (typically categorical) values the variable can take on. If the accurate and inaccurate match rates are similar across the values a given variable can take on (i.e., no association), then it can be said that for that variable, there is no specific value category that is performing significantly better/worse than the others. This facilitates the assessment of which demographic values are reported with higher accuracy by the TM data. To achieve this goal, results from the chi-squared test of independence are reported; however, due to the limitation that the chi-squared statistic is strongly influenced by sample size, Cramer's V is also reported (Cramer's V is a statistic that adjusts the chi-squared statistic using both sample size and number of cells in the contingency table). This adjustment allows Cramer's V to be comparable across contingency tables with different sample sizes and numbers of cells. Cramer's V ranges from 0 to 1, with higher values indicating high association.

Table 2.2 summarizes variable frequencies as well as measures of association across all variables and samples studied; note that the data used in this section is before matching on gender, age, and education had occurred so as to ensure that the results reported here are applicable to TM data in general (i.e., not biased by data processing). The final two columns of the table also present a direct comparison of the two survey datasets

to each other to provide some context for comparing the other distributions (i.e., how much congruence exists even between the same questions asked on two surveys of the same sample). As shown in the table, the chi-squared test of independence is significant for almost all variables, even in cases where Cramer's V is relatively small (see for example, household size). This is likely due to sample size effects, and accordingly, the Cramer's V statistic is primarily used for discussion here. Cohen's effect sizes (which are dependent on degrees of freedom) are used to select which effects based on Cramer's V are large enough to merit discussion (Cohen, 1988).

The Cramer's V statistic for age has a large effect size, and a closer look at the frequencies indicates that the TM dataset is doing a better job at reporting ages for individuals in higher age categories, which is intuitive given that these individuals likely have more established transactional histories, and accordingly their ages are likely to be better represented in TM databases. Tenure also has a large Cramer's V effect size, with the frequencies showing that TM data are doing a poor job in identifying renters, an intuitive finding given that: (1) renters are more likely to be lower-income individuals with fewer TM data records (and thus less accurate information); (2) renters tend to move more frequently than home owners, thus making it more difficult to maintain appropriate address information; (3) the apartment or unit number may be unavailable or incorrect for renters living in a multifamily dwelling at a given street address; and (4) renters may be living at rental properties that are single-family homes, making it difficult for TM data to accurately identify the tenure arrangement. Race also has large Cramer's V effect sizes, with the results showing that across TM and NHTS data, Asians and Native Americans are the most likely to be inaccurately represented, followed closely by African Americans. Thus, both

TM data and NHTS more accurately represent individuals who identify as White, a finding that may be attributable to Whites being more integrated into the financial/transactional fabric of U.S. society, and thus TM having more accurate records/sources of information for these individuals. It is likely also partially due to missing ethnicity being infilled by the TM data provider using aggregate data, with the dominant race at aggregate levels more likely to be White.

Dwelling type had the largest effect size across all variables studied, with the results showing that the TM dataset is much more likely to correctly identify individuals living in single-family homes. This is likely due to the same reasons discussed earlier for the tenure findings, and suggests that TM databases do not have reliable/accurate sources of information for individuals' living arrangements, particularly in cases where address details are less precise. Occupation also has a high effect size, with TM data being significantly more likely to inaccurately identify occupation type for those who are not in the professional, managerial, or technical category, although there are more inaccurate than accurate matches across all categories. NHTS is also more likely to differentially represent occupation type relative to GDOT survey responses for these categories.

Education is seen to have a medium to high effect size, with TM data being more likely to inaccurately identify individuals who have not completed high school and individuals with some college/technical qualifications. We note that education does present some difficult-to-interpret findings here, with individuals who have a completed high school degree or bachelor's degree being more likely to have correct matches, while individuals with some college/technical qualifications and those who have completed a graduate degree being less likely to have correct education records in the TM data. In

general, we would have expected that individuals with higher levels of education would have more sources of personal information (e.g., employment records) from which the education level can be gleaned, since in line with previous reasoning, they may have more established footprints in the TM database. This finding is discussed further in Section 2.6.3.

Marital status, household income, and household size have small effect sizes for the TM data comparisons in this study, and so deviations on these variables may be due to random fluctuations. However, it is interesting to note that three-person households are much more likely to have differences between the GDOT and NHTS surveys relative to the other household size categories, a finding that may be attributable to the one-year difference in survey administration for the GDOT and NHTS surveys. Further examination indicated that most of the incorrectly classified households in this category were two-person households in the NHTS survey that became three-person households in the GDOT survey, suggesting possible life stage changes like marriage or the birth of a child occurring in the (average) one-year gap between surveys.

To compare the findings from this study to the literature, it is seen that Kressner et al. (2014) used chi-squared tests of independence to examine patterns of association, and found no significant associations, with the primary exception of marital status. There was a higher occurrence of single individuals who had a correct match for marital status relative to married individuals, which Kressner et al. (2014) suggested may be because the TM database assumes that an individual is single until information is obtained that proves otherwise. However, the frequencies for marital status in the study presented here tell the inverse story, with TM data doing a better job of identifying marital status for those who are married. This difference in finding may be attributable to the particular population that

was sampled in the prior study or to differences in how the marital status variable was developed in the two separate TM databases. Kressner et al. (2014) also found that there were more households than expected whose targeted marketing data matched for the African American category, with fewer individuals who matched for the White category, but similarly we believe that this may be due to the distinctive study population sampled for that study, a proposition also suggested by the authors.

Table 2.2. Identifying patterns in accurate and inaccurate match rate distributions for sample pairs

Variable	SED characteristics	Frequency [a]					
		TM vs. GDOT [b] N = 3288 [c]		TM vs. NHTS [b] N = 5148 [c]		GDOT vs. NHTS [b] N = 1495 [c]	
		Accurate Matches	Inaccurate Matches	Accurate Matches	Inaccurate Matches	Accurate Matches	Inaccurate Matches
Gender	Male	1553	125	1861	275	661	10
	Female	1311	285	2594	416	803	15
	χ^2 statistic (df)	79.937 (1) ***		0.882 (1)		0.096 (1)	
	Cramer's V	0.156 (small [d])		0.013 (small)		0.008 (small)	
Age	18-24 years	13	20	35	68	8	2
	25-34 years	168	88	370	260	95	4
	35-44 years	265	65	524	213	127	11
	45-54 years	452	87	695	200	209	11
	55-64 years	667	115	952	228	346	21
	65+ years	1241	85	1447	152	626	29
	χ^2 statistic (df)	221.380 (NA [e]) ***		426.310 (5) ***		7.818 (NA [e])	
	Cramer's V	0.260 (large)		0.288 (large)		0.072 (small)	
Tenure	Owner	–	–	3384	224	–	–
	Renter	–	–	784	708	–	–
	χ^2 statistic (df)	–	–	1199.5 (1) ***		–	–
	Cramer's V	–	–	0.485 (med. to large)		–	–
Race	Asian/Pacific Islander	8	50	5	75	12	4
	Black/African American	362	197	801	418	243	0
	Native American	0	25	2	10	2	9
	White/Caucasian	2071	444	2900	637	1135	40
	χ^2 statistic (df)	305.240 (NA [e]) ***		381.600 (NA [e]) ***		220.230 (NA [e]) ***	
	Cramer's V	0.311 (large)		0.281 (med. to large)		0.390 (large)	
Marital status	Married	1446	459	–	–	–	–
	Single	665	389	–	–	–	–
	χ^2 statistic (df)	53.859 (1) ***		–		–	
	Cramer's V	0.135 (small)		–		–	

Table 2.2 cont'd

Dwelling type	Stand-alone house	1574	331	–	–	–	–
	Apartment/condo	61	993	–	–	–	–
	Mobile home	– [f]	– [f]	–	–	–	–
	Attached home/duplex/townhouse	0	318	–	–	–	–
	χ^2 statistic (df)	1953.200 (NA [e]) ***		–		–	
	Cramer's V	0.772 (large)		–		–	
Occupation	Professional, managerial, or technical	432	593	587	968	316	108
	Sales/service	24	278	31	601	54	44
	Manufacturing, construction, maintenance, or farming	21	57	42	228	22	4
	Clerical or administrative support	21	100	41	265	36	24
	χ^2 statistic (df)	139.910 (3) ***		302.550 (3) ***		20.107 (3) ***	
	Cramer's V	0.303 (large)		0.331 (large)		0.182 (large)	
Annual household income	Less than US $50,000	628	367	1604	833	485	45
	US $50-99,999	556	595	738	786	358	158
	More than US $100,000	502	511	510	601	327	75
	χ^2 statistic (df)	55.760 (2) ***		176.880 (2) ***		82.642 (2) ***	
	Cramer's V	0.133 (small)		0.187 (small)		0.239 (medium)	
Education	Some grade school/high school	0	74	0	177	24	11
	Completed high school or GED	155	199	293	592	155	12
	Some college/technical school	213	764	333	1224	407	71
	Bachelor's degree	443	546	500	751	368	58
	Completed graduate degree(s)	356	531	434	843	370	15
	χ^2 statistic (df)	176.870 (4) ***		202.360 (4) ***		46.669 (NA [e]) ***	
	Cramer's V	0.232 (med. to large)		0.198 (medium)		0.177 (medium)	

Table 2.2. cont'd

Household size	Single-person HH	331	592	713	1177	455	38
	Two-person HH	430	999	627	1296	550	83
	Three-person HH	90	344	187	446	98	75
	Four-person or larger HH	198	300	263	439	148	46
	χ^2 statistic (df)	47.836 (3) ***		21.124 (3) ***		132.59 (3) ***	
	Cramer's V	0.121 (small)		0.064 (small)		0.298 (large)	

***, **, * = significant at 1%, 5%,10%, respectively.

[a] Distributions examined before matching on gender, age, and education (i.e., before data processing) as described in Section 2.6.1.

[b] For the GDOT vs. TM and GDOT vs. NHTS distributional comparisons, the GDOT survey is used to inform the SED characteristics of the accurate and inaccurate matches for the contingency table. Similarly, for the NHTS vs. TM distributional comparison, the NHTS is used to inform the SED characteristics for the contingency table. This assumes that the survey data is "correct" relative to the TM data, which is not necessarily always true. Nevertheless, we have reason to believe that for most of the cases, survey data is likely to be more reliable relative to TM data. Furthermore, as the goal of the study is to study TM data relative to transport survey data, we believe that using the survey data sources to inform the SED tabulations for the contingency tables is appropriate.

[c] Counts do not add up to 100% or the total N because of noncomparable categories.

[d] Cohen's effect size classifications for Cramer's V are represented in parentheses following the Cramer's V value (Cohen, 1988).

[e] When the number of cases in a cell is small, a Monte Carlo procedure is used to calculate the test's p-value (Hope, 1968).

[f] The GDOT data did not have any individuals who reported living in mobile homes, so this category is not included in the distributional comparisons. Note however that the TM data *did* have 23 individuals who reported living in mobile homes.

2.6.3 Exploring biases for survey respondents more likely to be matched in TM databases

This section follows closely from the preceding sections, but refocuses the examination at the individual level as opposed to the variable level, examining the factors influencing individuals' propensities to be matched correctly in the TM database. Individuals are considered to have a correct match in the TM database if the survey record reflected the same gender, age within a +/- 4-year age tolerance, and education within a +/- 2-level tolerance with the returned TM record. Understanding which individuals may be better represented in TM databases facilitates an understanding of biases that can result when using TM data for transport applications. To assess these biases, we develop a binary logit model (Table 2.3) to predict whether a given respondent obtains a correct match in the TM database in terms of the gender, age, and education thresholds instituted during the

matching process (unmatched respondents also include individuals whose TM records were missing gender, age, and education, as these TM records could therefore not be checked relative to the respective survey record). For simplicity, this model is limited to the GDOT survey dataset (N = 3288; reduced to 3121 through the removal of missing values for this model); and the exogenous variables tested in the model include gender, age, race, education, occupation, household size, household income, marital status, and a measure of population density.

Gender, age, all levels of education, and race identification as African American or Caucasian are all significant predictors of the probability of receiving a correct match in the TM database used for this analysis. Women are less likely to be among those who have a correct match, an intuitive finding given that TM databases are largely derived from financial records and transactions which are often still dominated by males. Older individuals are also more likely to have a correct match, which may point to the increased probability of older individuals to have more established financial/transactional footprints. In the case of age, the inherent survey bias of the GDOT dataset toward older individuals is likely reflected in the small disparity between mean ages for the matched and unmatched records, and accordingly this suggests that there may be a greater difference between these means in a survey dataset that is more representative of all ages.

With regard to race, with Asian/Pacific Islander as the reference group, we see that Blacks/African Americans and Whites /Caucasians are significantly more likely to be among those who receive a correct match. However, as the incidences show, Blacks have a greater proportion of unmatched records than matched records (whereas the opposite is true for Whites), suggesting that while Blacks are more likely to be included in matched

records relative to Asians/Pacific Islanders, they are on the whole likely to be underrepresented in the TM database.

The model also indicates that, relative to individuals who have not completed high school, those with higher levels of education are more likely to have correct matches in the TM database, although those with graduate degrees are less likely to be matched relative to those who have completed some or all of their undergraduate education. This latter nuance, also shown in Section 2.6.2, may point to a higher proportion of foreigners among those with graduate degrees, relative to those with undergraduate degrees (National Academy of Sciences, National Academy of Engineering, & Institute of Medicine, 2007). Foreigners may be more likely to have incorrect TM records on several accounts; for example, individuals who have recently moved to a country are likely to have fewer records from both administrative and transactional sources. In addition, as before mentioned, gender misidentification may be higher for foreign names. Nonetheless, overall, the education findings suggest that TM databases may overrepresent more highly educated individuals, which is in line with the conceptual understanding that TM databases have more robust records for individuals with more financial assets and transactions associated with their names.

The findings in this section[1] support conventional intuition about the nature of TM databases, and along with the model findings in Table 2.3, serve to remind analysts

[1] For exploratory purposes, several TM variables were also tested in the model; however, since the TM variables for the non-matched individuals may not be correct at individual and/or household levels, we did not include these in the final model, but only mention them here. Two TM variables of interest that are significant are consumer prominence and technology adoption, with higher levels of both indicating increased likelihoods of having a correct record in the TM database. The consumer prominence indicator is a measure of how large the consumer footprint of the individual might be, while the tech adoption indicator is a measure of how likely a household may be to purchase new technologies at premium prices.

interested in using TM data that at least currently, there are certain demographics, notably groups such as women and certain ethnicities, that are disproportionately affected by underrepresentation in TM data.

Table 2.3. Binary logit model of whether a GDOT survey record is correctly matched to TM database

Variables [a]	**Coefficients**	**Variable Incidence (%)** [b]	
		2568 matched records	553 unmatched records
Constant	-1.593***	–	–
Gender (female)	-0.230*	47.08	54.97
Age	0.021***	60.11 [c]	54.97 [c]
Race	–	–	–
Reference group: Asian/Pacific Islander	–	1.32	4.34
Black/African American	0.889**	16.90	20.43
Native American	0.252	0.66	1.27
White/Caucasian	1.081***	81.11	73.96
Education	–	–	–
Reference group: Some grade /high school	–	1.79	4.52
Completed high school or equivalent	0.983***	11.06	10.67
Some college/technical school	1.204***	30.84	24.95
Bachelor's degree	1.233***	31.00	25.68
Completed graduate degree (s)	0.752**	25.31	34.18
Model attributes			
Number of observations	3121		
$\mathcal{L}(\mathbf{0})$	-2163.312		
$\mathcal{L}(\boldsymbol{c})$	-1457.822		
$\mathcal{L}(\hat{\boldsymbol{\beta}})$	-1401.567		
ρ^2 ($\mathcal{L}(\mathbf{0})$ base)	0.350		
Adjusted ρ^2 ($\mathcal{L}(\mathbf{0})$ base)	0.352		
ρ^2 ($\mathcal{L}(\boldsymbol{c})$ base)	0.039		

[a] The variables in this model are derived from the GDOT survey records for these respondents. As with all data sources, the GDOT survey may have its own implicit survey/nonresponse biases that may influence these numbers.
[b] Variable incidence represents the percentage of matched and unmatched records falling into the respective variable categories; for example, 46.85% of the matched records are females, while 56.67% of the unmatched records are females. Again, the GDOT survey was used to obtain the values for these variable incidences.
[c] As age is a continuous variable in the model, the mean ages for the matched and unmatched records are reported here in place of the incidence. Thus, note the sample bias toward older ages, even among unmatched records but especially among matched records.

2.7 Discussion

Using various validation methods, it has now been shown that TM data are able to provide accurate information (relative to self-reported data) for some variables and

populations, while underrepresenting others. This is not unexpected, given that all data sources, active and passive, will inevitably suffer from unique biases and shortcomings. In fact, this serves to reinforce the earlier suggestions that it is critical for researchers working with new data sources to first validate novel data using an array of methods, and preferably, to also have these data validated by differing teams of researchers. Without undertaking thorough validation investigations, biases present in various datasets may be unknowingly integrated into decision-making processes and affect key transport outcomes like equity and wellbeing. While it is outside the scope of this document to provide an extensive discussion on approaches that can be used to address dataset biases (for example: Cahan, Hernandex-Boussard, Thadaney-Israni, & Rubin, 2019), the aim here is to provide a brief recap of the validation exercises, and to provide recommendations for methods that may be useful in the specific context of the TM data being examined in this study.

In Section 2.6.1, it was seen that TM data is able to provide accurate data on several key variables (gender, age, tenure, and race) for 75% or more of individuals in the two survey samples studied. It is an opportune time to emphasize a point first made by Kressner et al. (2014), that even the variables that were found to have the lowest accuracy rates (~31-34%), indicate that with TM data, it may be possible to accurately predict these variables for at least a third of the population at a significantly lower cost than it would take to acquire these variables using surveys. In Section 2.6.2, distributions of accurate and inaccurate matches across all variable values were examined to provide an understanding of how specific categories of each variable are performing. This investigation showed that age, race, dwelling type, occupation, and education perform differently across categories. This means that it may be especially important to realize that TM data may be providing

incorrect information at a higher rate for certain individuals; for example: younger individuals, renters, minorities, etc. In Section 2.6.3, the biases present for those who were considered to have a correct record match in the TM database were explored, and indicated that at an individual level, women, minorities, younger individuals, and those with lower levels of education are less likely to have a correct record in the database.

Practitioners seeking to address biases such as those described here may: (1) seek to augment the data source in question with additional records/cases from other data streams that may be more representative of specific populations (i.e., data fusion); (2) develop algorithms/models to impute variable values for segments of the population that have increased probability of having incorrect values; (3) identify the need to develop weights that can adjust the sample for the variables on which biases have been identified (Solon, Haider, & Wooldridge, 2015); and (4) interpret results within the lens of the biases that may exist, ensuring that the proper caveats are applied when making policy recommendations. These approaches represent some of the possible solutions that could be applied to address the TM data biases identified in the preceding section. However, there are certainly other approaches, and all transportation researchers and practitioners who work with user-centered data should make it a priority to explore the methods and approaches that can be used to address dataset biases.

2.8 Summary and conclusions

Given the "growing resistance among U.S. householders to surveys in general" (PTV NuStats, 2011, p. 43), it is increasingly important to examine additional sources of data that can be used to supplement transport modeling needs. This chapter made the case

that targeted marketing (TM) data are ripe for integration into transportation applications, beginning with a detailed look at the benefits and challenges of using TM data (Section 2.3). The presented typology illustrated that TM data can be useful to a range of transportation applications and research, allowing for improved transportation models and innovative approaches that could reduce analysts' reliance on traditional transportation data sources (Section 2.4). Based on the experience integrating TM data with two transportation surveys (NHTS and a GDOT-funded survey), a framework of the TM data enrichment process was detailed (Section 4), providing a case study of the process for analysts who may wish to pursue similar TM data integration and enrichment (Appendix B.2).

The resultant integrated datasets were used to demonstrate that TM data match gender, age, tenure, race, marital status, and household size at rates of 70% or greater relative to self-reported survey data (Section 2.6.1). However, it was seen that TM data exhibit differential accuracy across some variable categories; for example, the database does a poor job correctly identifying tenure and dwelling type for renters and those not living in single-family homes (Section 2.6.2). This may suggest that transportation professionals who use TM data in the future may need to impute or otherwise supplement data for demographic categories that tend to be inaccurately reported in TM databases. Additionally, an examination of TM biases revealed that men, older and better-educated individuals, African Americans, and Caucasians are more likely to have correct records in TM databases (Section 2.6.3). These are comparable though not identical to typical HHTS respondent biases, suggesting that similar approaches taken to address biases in transportation survey data may need to be applied here (Section 2.7).

There are numerous avenues of future work that can be pursued in the aim to better understand the potential benefits of TM data in transportation. Notably, practitioners may be interested in better understanding the veracity of travel behavior variables that are present in TM databases, and thereafter, to investigate the use of such variables within forecasting efforts. To date, the authors are aware of only one paper that has sought to examine the veracity of travel behavior variables present in TM data, and that work has yielded promising results that certainly call for the further investigation of such variables by all transportation professionals (Lovelace et al., 2016). In addition, it will certainly be critical for the transportation community to have various teams of researchers investigate the applications and research directions proposed in the typology described in Section 2.4, as currently only a handful of studies thus far have tested TM data in similar applications or contexts. Of special interest will be methods for integrating and fusing TM data with other active and passive data sources, as this approach will aid in overcoming biases present across the various data sources while creating an enriched dataset that can facilitate novel analyses and insights.

However, while there is clearly significant potential in the use of TM data in transport applications, there remain challenges hindering the wide-scale application and integration of these data for modeling purposes in the transport domain – challenges that could intensify as we move through a period of increasing privacy regulations. Both as engineers and as private citizens, it is best to pursue TM data research and practice opportunities that will protect individuals' privacy while allowing for societal gains. It will be increasingly important for professionals to work with policymakers to strike such a balance, particularly in light of the growing need to supplement traditional data sources

with various passively collected data sources, all of which are subject to the same privacy regulations discussed in this chapter. In closing, it is hoped that this resource will encourage transportation professionals to further explore the benefits of targeted marketing data for moving transportation research and practice forward, while encouraging the contribution of new perspectives on approaches and methods that can be used to address some of the many challenges inherent not only in TM data, but third-party, passive, big data sources at large.

CHAPTER 3. EXPANDING SURVEY DATASETS THROUGH PREDICTIVE TRANSFER

Surveys are a key data source in transportation, as well as in an array of other disciplines; however, as discussed in Chapter 1, they are facing mounting challenges that may threaten their viability and long-term sustainability for providing critical information on which analysts have long depended to forecast future trends and make policy decisions. Approaches for integrating and/or enriching survey datasets with other surveys can expand the information available for forecasting efforts without generating additional burden upon survey respondents. This chapter provides:

1. an overview of methods that have been used in transportation for enriching survey datasets;
2. a detailed presentation of the transfer learning approach, a method that allows analysts to use advanced algorithms and passive, big datasets to improve the quality and thus, value, of survey enrichment efforts; and
3. a step-by-step application of the transfer learning approach to bring attitudinal variables into the NHTS (in Chapter 5, the results of the application are externally validated).

By demonstrating a systematic approach to investigating a scalable, advanced survey enrichment approach, the intent is to provide a prototype that can be used by analysts seeking to enrich survey datasets with values from other sources.

The work detailed in this chapter is from the following paper, which is currently in preparation:

Shaw, F. A., Wang, X., Mokhtarian, P. & Watkins, K. (paper in preparation, available upon request from authors). A framework for enriching survey datasets using big data and machine learning, with an application for transferring attitudinal variables across transport surveys.

3.1 Abstract

Declining survey response rates make it increasingly critical for survey designers across disciplines to utilize mechanisms that facilitate timesaving and reduce the burden on the part of respondents. In practice, this often means that questionnaires are shortened, yielding increased response rates but reduced information/variables available for modeling and forecasting purposes. Here, this challenge is addressed by making use of data driven approaches like machine learning, within the context of the rapidly growing big data landscape, to develop and apply a predictive transfer learning-based framework for enriching surveys with information from other survey datasets, thereby expanding the amount of information available for use. The framework is demonstrated by applying it to supplement and enrich the U.S. National Household Travel Survey (NHTS) with psychometric data (e.g., attitudes, preferences), which have been shown in the literature to have the ability to explain and predict behaviors, but which are often not captured on household travel surveys. Using the framework presented, it is shown that it is possible to train algorithms that can explain up to 25% of the variance in observed attitudes, yielding correlations of up to 0.5 between observed and predicted attitudinal variables. Applications of the framework presented in this chapter have the potential to improve travel demand forecasting and behavioral predictions; and, even more broadly, may be used to enrich

other large-scale behavior-based surveys with external variables, thereby providing more diverse and robust data streams for use in an array of modeling efforts.

Keywords: data fusion; imputation; machine learning; household travel survey; transportation survey; travel demand modeling; consumer data; targeted marketing data; attitudes; attitudinal constructs; psychometric variables

3.2 Introduction

Evidence confirms that survey response rates have been falling steadily for over half a century, and researchers agree that the field may be converging upon a critical point at which the validity of survey findings is increasingly called into question (Lohr & Raghunathan, 2017; National Research Council, 2013; PTV NuStats, 2011). Theories of survey response find that respondents fail to complete surveys for a plethora of reasons, critical among them, increased concerns over intrusions on time and privacy (Goyder, Boyer, & Martinelli, 2006). As demands upon individuals' time continue to grow, the Social Exchange theory of survey response explains that perceived benefit for the "cost" of response time is decreasing (Dillman, Smyth, & Christian, 2014). This is supported by empirical evidence showing that collective attention span is decreasing due to an overload of content that exhausts attention resources (Lorenz-Spreen, Mønsted, Hövel, and Lehmann, 2019). Accordingly, it is increasingly important to attend to efforts that facilitate timesaving and reduce the burden on the part of respondents. This has resulted in widespread efforts by survey designers to reduce the lengths of survey questionnaires, thereby improving response and completion rates but simultaneously reducing the amount of information obtained.

The implications of reducing survey length are particularly pertinent within fields like transportation, where engineers and planners depend upon long-form travel diary and survey data to forecast evolving infrastructure needs. As mentioned, the poor performance of travel demand forecasting models is well documented (Bain, 2009; Hartgen, 2013; Nicolaisen & Driscoll, 2014; Parthasarathi & Levinson, 2010; Voulgaris, 2019; Welde & Odeck, 2011), with current models often operating at less than 10% explanatory power and requiring subjective alterations to improve performance. Such poor model performance is partially attributable to the lack of diverse variables such as attitudes, preferences, perceptions, social and personal values, and other such system user traits (i.e., psychometric data) available for use within forecasting models. Furthermore, the data/variables needed to answer complex research questions are seldom available through a single survey dataset (Sivakumar & Polak, 2009). With the increasing need to shorten questionnaires, this lack of availability of diverse variables promises to be a growing challenge.

Addressing this challenge will necessitate a broad range of approaches centered around improving data quality and richness. Recent efforts have typically focused on the use of novel non-survey-based data sources to support transportation modeling (see for example, Chapter 2 of this thesis for a discussion on this subject). However, forecasting travel behavior still depends on household and individual-level survey data, due largely to the user-verified, self-reported nature of survey responses, alongside their ability to obtain domain-specific data that often is not (easily) available through other data streams. Accordingly, in this chapter, the focus is on developing a flexible framework for expanding the data available from surveys by enriching/integrating survey datasets (“recipient

surveys") with survey variables outside of the original survey domain (i.e., from "donor" surveys). In order to maximize all tools available, the framework uses novel, big data sources alongside data driven machine learning (ML) algorithms; but it is also shown that the essence of the data transfer framework can be applied even in the absence of these tools (i.e., within a simpler context).

The remainder of this chapter is organized as follows. It begins by reviewing survey data enrichment literature, with an emphasis on methods and applications in the transport domain (Section 3.3). Building on this synthesis, a methodological framework for survey enrichment is presented that represents a core contribution of this chapter (Section 3.4). Next, the transfer learning framework is applied to transfer attitudinal variables from a small, variable-rich research-oriented survey (GDOT Survey) into a larger, nationwide travel survey (NHTS Survey; Section 3.5). The chapter closes with a discussion of key takeaways from the framework and application (Section 3.6) that is intended to be helpful to researchers and practitioners in all domains. Over time, the efforts initialized by this work are intended to provide more diverse and robust data streams for use in modeling and forecasting efforts.

3.3 Survey data enrichment methods in the literature

Given that survey data have long been among the most critical of data sources for transport modeling, it is unsurprising that a plethora of literature in transportation has applied various methods for enriching and expanding the information obtained from survey datasets. In general, these enrichment efforts have used the following approaches: (1) long-form, repeated or follow-up survey sampling; (2) data augmentation; and (3) statistical

methods for integrating survey datasets. Figure 3.1 summarizes the primary methods used in transportation for *directly* expanding (i.e., bringing new variables) survey datasets. In addition to these methods, there are other forms of survey data enrichment that allow for the *use* of data from other surveys, albeit without directly importing new variables, e.g., parameter estimation using information from external datasets (Lohr and Raghunathan, 2017). The overview shown in Figure 3.1 emphasizes survey data enrichment methods that bring new variables entirely into the recipient datasets, as this is the form of enrichment that is being studied within this chapter. For a broader and more exhaustive typology of data enrichment methods, the reader is referred to Zheng (2015). Note that the term data fusion has been intentionally omitted from this discussion due to its lack of clear definition and rampant conflation with numerous methods and purposes across fields (D'Orazio, Di Zio, & Scanu, 2006; Malokin, 2019; Tsamardinos, Triantafillou, & Lagani, 2012).

Before beginning this discussion of survey data enrichment in transportation, readers are reminded that all data enrichment procedures, such as the ones that are discussed within this chapter, can engender privacy concerns due to involving large amounts of information at an indiviudal or household scale. However, since these procedures inevitably introduce various errors and biases, the information gathered is – comfortingly, from this light – not always accurate. Nonetheless, analysts using linked/enriched databases should ensure that consent and privacy regulations are followed, and that all team members are trained in the ethical handling and usage of the resulting datasets.

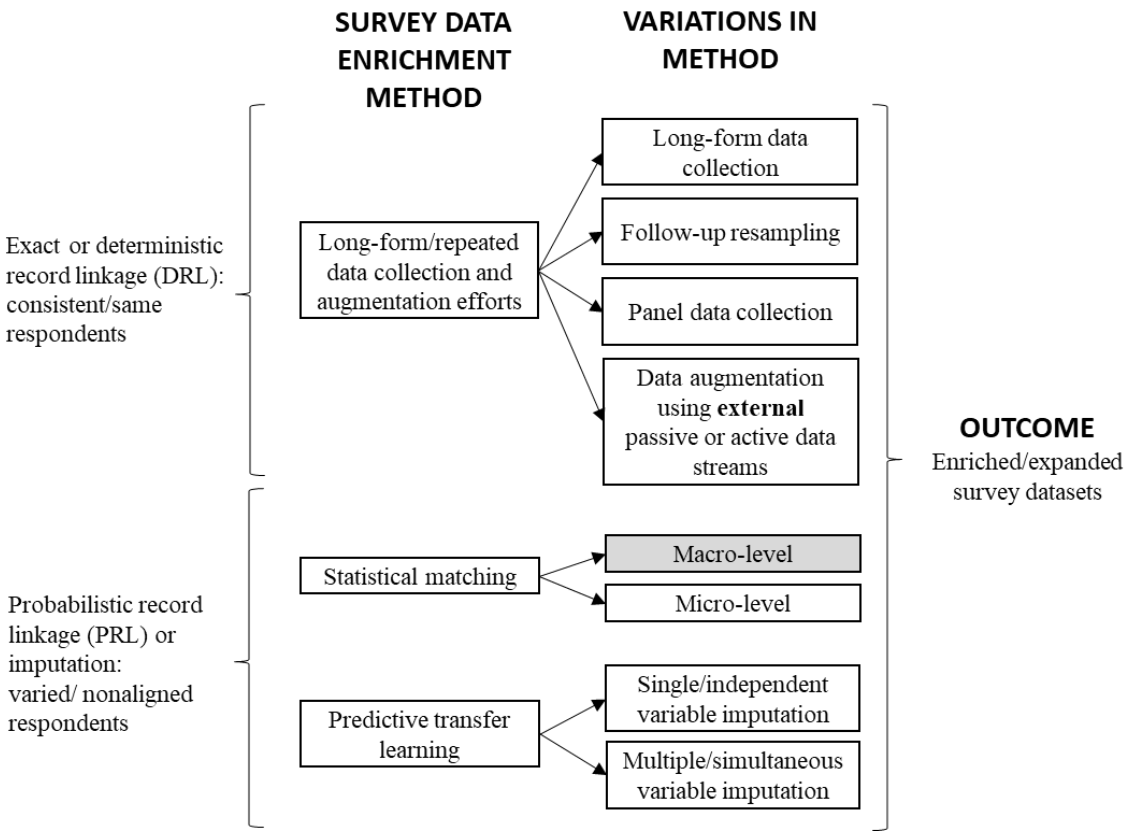

Figure 3.1. Overview of survey data enrichment methods for transportation

* This approach is in grey as it does not directly bring new variables into the receiving dataset.

3.3.1 Exact or deterministic record linkage across consistent respondents

The most straightforward method for obtaining a wide array of variables for the *same* respondents entails: (1) long-form or repeated survey data collection efforts (i.e., through active data collection) or (2) the augmentation of existing survey datasets with external variables (i.e., data available through active or passive data collection that occurs independently of the recipient survey) that can be accurately attributed to the appropriate respondents. As shown in Figure 3.1, these approaches are classified as exact or deterministic record linkage (DRL; terminology from Newcombe et al., 1959; Winkler, 1999; Lohr & Raghunathan, 2017), as they allow for the linkage of records using a set of characteristics that are believed to uniquely identify individuals (i.e., the same entities);

however, this does not mean that there are no linkage errors as even reported names can vary across datasets (e.g., Bob, Robert, Rob). Traditionally, this group/class of approaches has been the most relied upon by transport survey designers as they do not require knowledge or use of various statistical methodologies/programs. Unfortunately, DRL methods face significant and growing challenges with regards to response rates, survey implementation procedures, and recruitment/data acquisition costs (Hössinger et al., 2020; Lohr & Raghunathan, 2017).

Firstly, obtaining long-form, active datasets involve the direct collection of a large amount of information from the same respondents, either simultaneously, via a follow-up survey, or with the use of a longitudinal panel. To clarify, the term "follow-up survey" here refers to respondents who are recruited from preceding surveys; that is, respondents who completed an initial survey and agreed to be contacted again for future surveys. This differs from longitudinal panels, as follow-up surveys tend to be conducted by independent research teams with different questionnaires, while the goal of longitudinal panels is typically (though, not always) to collect the same variables from the same respondents across time. As mentioned, there is an array of challenges associated with these approaches. Notably, longer, more detailed surveys have lower response rates while follow-up and panel surveys can yield high attrition rates and unrepresentative samples (Adriaan & Jacco, 2009; Amarov & Renatel, 2013; Couper, 2007; Wang, Shaw, Mokhtarian, & Watkins, in preparation). These challenges can make it extremely difficult and/or cost prohibitive to obtain reliable and varied datasets across the same individuals using survey sampling methods.

Meanwhile, the fourth variation of this method (as shown in Figure 3.1) – data augmentation using external (active or passive) datasets for respondents on whom survey data is available – holds its own challenges. Firstly, it is all but impossible to find multiple *active* datasets on the same respondent, unless that respondent is part of a panel or is recruited through follow-up sampling as discussed in the prior approaches. With regard to obtaining external passive datasets, there are often privacy constraints that make it difficult to obtain passive data on a *specific* individual, making it necessary to rely on the use of statistical methods such as the next two methods discussed (i.e., statistical matching and predictive transfer learning). When it *is* possible to purchase external, passive datasets for specific survey respondents (as in the case of Chapter 2), there are further challenges involved in data integration and use. These can range from incomplete external datasets (i.e., extensive missing data) to inaccurate/out-of-date variable values (Shaw et al., 2020).

Methodologically, the DRL class of approaches, which entails obtaining rich data across the same respondents, is attractive because it bypasses many of the statistical assumptions and errors present when merging disparate datasets as is the case for the next two survey data enrichment methods discussed (Hössinger et al., 2020). However, in light of the significant and growing challenges associated with these DRL approaches, it is clear that analysts may have to increasingly turn to probabilistic record linkage (PRL) approaches such as those discussed next for reliably obtaining rich, diverse datasets.

3.3.2 Statistical matching

The most widely used survey integration method in transportation is statistical matching, a stochastic/probabilistic procedure that merges disparate datasets using

distance-based measures that capture the similarity between cases based on one or more common variables that are present in all the datasets being merged. Like the survey enrichment methods discussed in the previous section, all variables being transferred in the statistical matching process are still observed, but not observed from the same individual. Instead, these enriched/transferred variables are borrowed from a different individual who share some similarities based on selected common variables available between datasets. As shown in Figure 3.1, statistical matching has two primary approaches. The first is the macro approach, and entails using the data present in both datasets to derive estimates of parameters of the joint distribution of the unique and common variables (e.g., correlation coefficients or contingency tables). This approach is shown in grey in Figure 3.1 as it does not directly bring new variables into the receiving dataset. The second statistical matching approach is termed the micro approach, and involves the creation of a new synthetic dataset that comprises all unique and common variables across all datasets being merged (D'Orazio, 2017; Konduri, Astroza, Sana, Pendyala, & Jara-Díaz, 2011; Lohr & Raghunathan, 2017; Müller & Axhausen, 2014).

Statistical matching is often applied by transportation researchers to answer research questions that may require data from multiple surveys. Here, some examples of such applications are provided. For example, Sivakumar & Polak (2009) sought to study the relationship between leisure activity participation and household technology holdings; and to do so, needed to combine the UK National Travel Survey (NTS) with the UK Time Use Survey (TUS). NTS provided transport related data regarding out-of-home activity locations or the choice of mode when travelling to out-of-home activities, while the TUS provided details regarding technology holdings and time spent on in-home and out-of-

home leisure activities. The study used both statistical matching and transfer learning methods, with the statistical learning approach using ad-hoc cluster sampling to integrate the surveys. In ad-hoc cluster sampling, clusters were created in the TUS datasets using variables common to both NTS and TUS. For each respondent in NTS, a random TUS respondent from the same cluster is selected, and their technology variables are assigned to the receiving NTS respondent. Similarly, Pawlak et al. (2013) sought to study the relationship between digital behavior and physical mobility and used the k-nearest neighbor method to match cases between travel diary and lifestyle datasets. In addition to such applications, statistical matching has also been used to supplement population generation methods in transportation, i.e., to bring additional variables into synthetic populations (Müller & Axhausen, 2014).

Statistical matching has several benefits. Firstly, it is a non-parametric approach which means that the resulting dataset is not affected by distributional assumptions of an algorithm. In addition, the method allows for the easy transfer of multiple variables simultaneously as the entire record for each respondent can be transferred from the donor sample. As a result, statistical matching is efficient in keeping "covariance structure and avoiding incoherencies" (Pawlak et al., 2013, p. 4; Saporta, 2002, p. 471).

Among the most substantial challenges facing statistical matching are the assumptions that must be made about the comparability of the datasets being integrated, as well as the conditional independence assumption which requires the common variables to explain all of the association between unique variables in the donor and recipient datasets (or, put another way, the occurrence of the unique variables must be independent of each other, conditioned on the common variables; D'Orazio et al., 2006). Both of these

assumptions are very difficult to meet, with the latter being almost impossible to verify (due to the absence of a dataset that includes both sets of unique variables as well as the common variables). In addition, another significant challenge lies in the relevance of the common variables being used to join the datasets relative to the unique variables. For example, if the common variables being used to join datasets are age and gender and one of the unique variables (in one of the datasets) is attitude toward working, it is not justifiable to assume that all women in a certain age range would have similar attitudes toward working, as variables like personality and education may also play important roles in determining work-oriented attitudes. Further exacerbating this challenge is the fact that it is not possible to judge the performance of the matching process since typical goodness-of-fit measures cannot be obtained; however, to counter this, note that internal validation can provide avenues for evaluating the differences in transferred variable distributions and relationships with other variables between the fused and donor datasets. Lastly, statistical matching may underestimate the variability present in the data (e.g., by using some donor cases more than others) which can result in increased Type 1 errors, a challenge that may be mitigated by the introduction of randomness in the imputation (Pawlak et al., 2013; Rubin, 1987).

3.3.3 Predictive transfer learning

Transfer learning is, to our knowledge, the enrichment method that is least explicitly discussed in the *transport* literature and represents the underlying approach that informs the framework presented and applied in this chapter (Section 3.4). At its core, this method involves the *predictive* transfer or predictive imputation of variables from one dataset (donor dataset) to another (recipient dataset), using the common variables present

in both datasets as the explanatory variables or features. This method is commonly applied within the domain of computer science, and particularly, its subdomains of artificial intelligence and machine learning (Pan & Yang, 2010; Tsamardinos et al., 2012; Zheng, 2015). Further, across a broad swath of disciplines, the transfer learning method has been used as a missing data imputation method *within* datasets. Thus, the use of transfer learning is not new; however, to our knowledge, there has not been a systematic framework presented for its application within transportation or urban planning, which prevents critical cross domain linkages between transportation and computer science that could benefit data enrichment efforts in transportation. Accordingly, this section focuses on a summary of the method and its applications as detailed in the literature, with a more technical discussion of transfer learning presented in Section 3.4.

3.3.3.1 Terminology

To help establish clarity in terminology, some of the terms used in Figure 3.1 and/or the literature are briefly discussed here. The term "transfer learning" originates from the computer science field and is an umbrella term for data fusion that involves the transfer of knowledge from a source domain to a target domain. Thus, the technical definition of transfer learning can theoretically encompass both deterministic record linkage and statistical matching as well as the application under discussion. However, in the case of this chapter – and what is hoped will be a precedent for its use in our discipline – the term is used to explicitly refer to a sub-instance of transfer learning that is more common and applicable within transportation; namely, "heterogeneous transfer learning" which is the transfer of information amongst datasets with different feature spaces (Zheng, 2015). To help cement the more specific definition intended in this field, the term "predictive" is

included prior to “transfer learning”. In transportation, it is also seen that Pawlak, Polak, and Sivakumar (2013) use the term “implicit imputation” to refer to statistical matching and the term “parametric, explicit imputation” to refer to transfer learning that is executed with the use of parametric models (we note that transfer learning can also occur with nonparametric models as shown in Section 3.5 of this chapter). As shown in Figure 3.1, both statistical matching and predictive transfer learning are considered data imputation approaches, and accordingly both of these approaches are grouped within the probabilistic record linkage or imputation domain.

3.3.3.2 Variations in method

There are two main avenues for varying the predictive transfer learning method (see Figure 3.1) that are relevant within the transport field (with many other avenues detailed in computer science (Zheng, 2015)). These include the transfer of single variables independently, or the transfer of multiple variables simultaneously. In this chapter, the focus is on the independent/single variable transfer learning process, examining parameters that can be varied within this process. The developed framework and parameter changes for single variable transfer can be applied by transportation analysts seeking to use simultaneous or multiple variable transfer learning. The two major groups of parameters that can be varied when developing a transfer learning framework include the common variables and algorithms used to transfer the variables. In this chapter, both basic linear models as well as ML algorithms are examined. Transfer learning algorithms used in the transport literature include regression and choice models, Rubin’s multiple imputation, and Bayesian conditional probability models (Eisenmann & Kuhnimhof, 2018; Sivakumar &

Polak, 2009; Sivakumar & Polak, 2013). Common variables and transfer algorithms are discussed in greater detail in Sections 3.4 and 3.5.

3.3.3.3 Applications in transportation

As in the case of statistical matching, transfer learning is typically used in transportation applications when the information needed to address a research question is not easily available from one data source. However, within the transport literature, the transfer learning method is often not the focus of the research and so clear details regarding the data enrichment process are seldom presented. As noted in Section 3.3.2, Sivakumar & Polak (2009) used both statistical matching and transfer learning methods, with the transfer learning approach using Bayesian models to integrate the NTS and TUS surveys. In another example of transfer learning, Eisenmann and Kuhnimhof (2018) aimed to study the relationship between the costs of car ownership and travel behavior, and to do so used linear regression models to bring costs from the German vehicle cost database into the German national travel survey. However, in neither of these studies were consistent terms used for the transfer process, a problem that occurs throughout transportation and which can make it difficult to locate papers in the literature that have used variants of predictive imputation methods to integrate datasets in transportation. A more exhaustive and systematic review paper on this subject would be a significant contribution to the field.

3.3.3.4 Benefits and disadvantages

A benefit of the transfer learning approach is that various goodness-of-fit statistics can be easily obtained to measure the performance of the transfer models, a benefit that is not possible with the statistical matching approach (Pawlak et al., 2013). Further, as will

be shown, it is possible to improve the transfer learning process with the use of augmented common variables for the recipient and donor datasets (derived using external datasets); meanwhile, increasing the number of common variables simultaneously used in statistical matching can often lead to reductions in number of matchable cases (as more common variables may increase the likelihood of nonmatched cases). This observation is logical and was also confirmed through internal validation exercises by our research team, during which we found that transfer learning outperforms statistical matching in the application that is shown in this chapter (Section 3.5). This work is forthcoming and will serve as a complement to the work detailed within this chapter.

There are several statistical drawbacks that can arise when using transfer learning to enrich survey datasets. Firstly, the donor and recipient datasets should be from the same population; however, even in this case, different sampling and measurement errors and uncertainties are present. In addition, as in the case of statistical matching, the common variables present between the disparate datasets are of course imperfect predictors for the variables being transferred (i.e., they cannot explain all the variance present in the transfer variable), a fact that can result in poor transfer performance and/or large errors in the imputed variables. Further, the error terms for the donor dataset, recipient dataset, and imputed variables may have different scales/distributions, an outcome that can potentially affect future modeling efforts (Sivakumar & Polak, 2013). This latter limitation is often one reason that multiple imputation (MI) is recommended in such cases (i.e., because the MI estimator seeks to quantify the effects of measurement uncertainties). In addition, the use of parametric algorithms for the transfer process (e.g., linear regression) can result in undesirable distributions or other statistical phenomena present when using the merged

dataset. Lastly, the use of transfer learning independently for a set of variables (i.e., imputing one variable at a time) can result in between-variable inconsistencies (Pawlak et al., 2013), which encourages the further exploration of multiple/simultaneous variable transfer/imputation.

3.4 Transfer-learning framework

The following subsections provide a technical overview of the transfer-learning based framework for transferring variables across datasets, followed by a practical overview of the components of the framework. As explained, this framework can be used to enrich survey datasets with variables from other surveys, thereby eliminating the need to ask these questions on the recipient survey itself. In addition, this framework can facilitate the integration of difficult-to-capture variables, as well as out-of-domain variables, thereby expanding the amount of information available for modeling/forecasting applications.

3.4.1 Overview of methodology

Figure 3.2 summarizes the methodological process of this study using the framework developed in Pan & Yang (2010), and first applied within a similar research context by Malokin (2019). In the terminology used here, a dataset, D, comprises a full p-dimensional space $\mathcal{X}$ (subspaces of which are represented by X) and a full q-dimensional space $\mathcal{Y}$ (subspaces of which are represented by Y). Specifically, the donor dataset is represented by D_D, which is defined as the set of output variables that are of interest to be transferred $\mathcal{Y}_D$, plus the set of remaining variables in the donor dataset, $\mathcal{X}_D$.The donor dataset input variables to be used as part of the algorithm training process represent the

common variables between donor and recipient datasets (X'_D) and thus constitute a subset of the total available variables with the exception of the transfer/output variables ($\mathcal{X}_D$). In parallel, the recipient data source, D_R, is defined holistically as having a set of total available variables $\mathcal{X}_R$, and the recipient dataset input variables that are common to the donor and recipient datasets are denoted as X'_R. Additional variables, X''_D and X''_R, represent variables unique to the donor and recipient dataset respectively (i.e., not present in the other dataset), but which are not used in the transfer process. Given these definitions, a learning function $f(\cdot)$ is developed that learns to predict Y_D based on X'_D, and then this function is applied to X'_R to predict $\hat{Y}_R$. Thus, $Y_D = f_D(X'_D) + \varepsilon_D$, and $\hat{Y}_R = f_D(X'_R)$, where the learning function f_D is invariant between the donor and recipient domains.

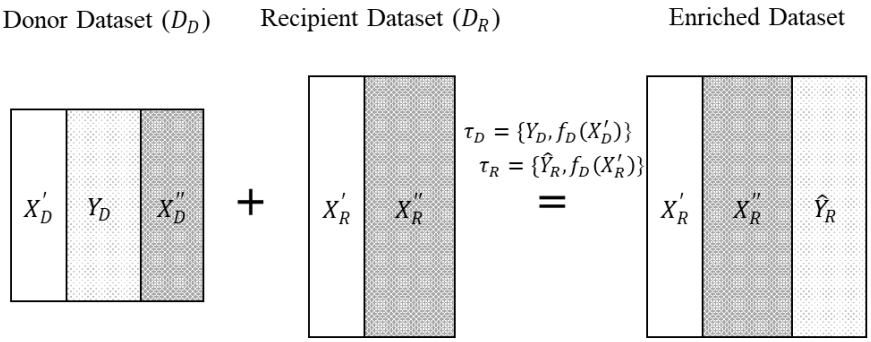

Figure 3.2. Methodological overview of study process
Source: Derived from van der Putten et al. (2002) and Malokin (2019)

Within the transfer learning framework, there are basic assumptions or requirements that are used to assess transfer compatibility between the donor and recipient datasets. Firstly, the datasets are advised to have similar spatial and temporal characteristics in an effort to minimize their differences on unobserved attributes. Next, the common variables between the donor and recipient datasets (X'_D and X'_R) are expected to have consistent definitions, measurements, and distributions. Accordingly, common variables

should be aligned with regard to variable categories and units of measurement. Procedures to adjust the distributions (e.g., weighting and sampling) may also be examined, but ultimately, the intent is to make the best effort to address dataset differences while acknowledging that the act of fusing disparate data sources inherently means that there will be both observable and unobservable sources of differences between datasets, some of which will be addressable while others are not. The ultimate objective is to assess whether or not the transferred data, in all of its imperfection, is still ultimately useful in the intended application – whether that is by bringing additional insight or by improving forecasting efforts, or both.

3.4.2 Components of transfer process

As shown in the preceding section, the theory behind transfer learning is straightforward; however, the execution can be complicated, depending on the datasets being used and parameters that the analyst chooses to vary. In some instances, a simple and clear path forward may be preferred; while in other cases, the analyst may wish to examine performance differences resulting from a wide combination of different parameters. This chapter provides a detailed overview of many of the various parameters that can be adjusted should the analysts have the available data and/or resources to do so. Figure 3.3 summarizes the three primary components of the transfer process and provides examples of adjustable parameters for each of the components. These components are: (1) the variables to be transferred; (2) the algorithms (also known as functions or models) that are trained to predict the transfer variables; and (3) the features (or explanatory variables) that represent inputs into the transfer algorithm. Each of these three components are discussed independently in the following subsections.

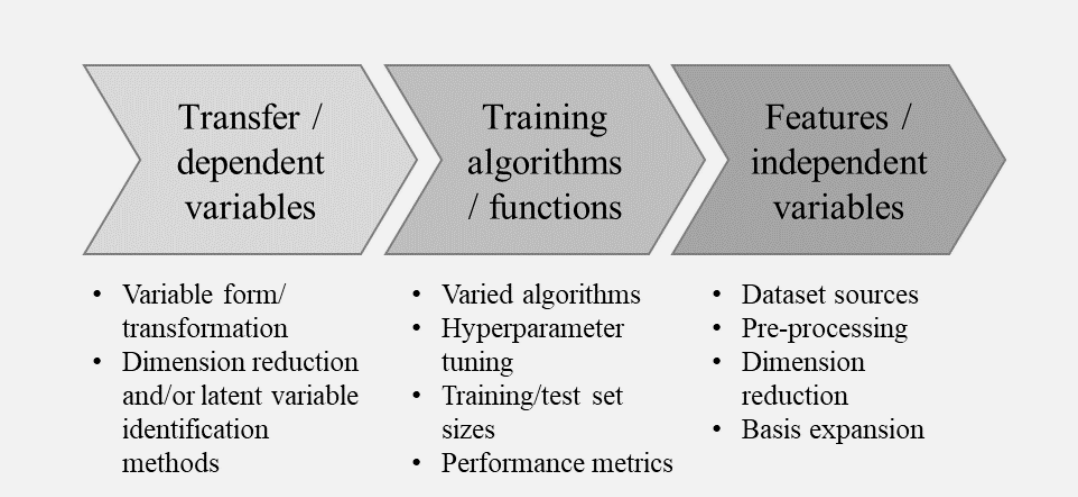

Figure 3.3. Overview of components and sample parameters in transfer process

3.4.2.1 Transfer/dependent variables

As the core goal of the transfer methodology is to transfer variables of interest from one dataset to another, it stands to reason that the transfer variables are the most important component that should be examined in detail (note that for consistency, throughout the remainder of this chapter, the term "transfer variable" is used when referring to variables that are to be transferred across datasets). Depending on the transfer variable type(s), analysts may alter the variable so that the variable form that is most accurately transferred is the one selected for use during the process. For example, when transferring categorical variables, combining or separating various categories may result in improvements in transfer accuracy. Alternatively, depending on the algorithm being used to transfer the variable, it may be that linear/nonlinear transformations of the transfer variables result in improvements in transfer accuracy. Thus, as can be seen, determining the best version of a transfer variable is analogous to the approach taken when modeling any dependent variable of interest.

In addition, analysts may choose to use dimension reduction approaches to transfer linear/non-linear combinations of transfer variables. Similarly, if the variables are psychometric in nature, latent variable identification methods may be used to transform indicator variables into latent, continuous variables. In these cases, various dimension reduction or variable identification methods can be tested; examples include cluster analysis, principal components analysis (PCA), exploratory factor analysis (EFA), and confirmatory factor analysis (CFA). Further, different numbers of constructs or clusters could be extracted to represent the final variables to be transferred. Alternatively, analysts may choose to transfer the discrete indicator variables first and then apply the reduction or identification methods after the transfer. Thus, it is seen that many different forms of the transfer variables can be examined to ensure that the best path possible is chosen during the process.

3.4.2.2 Features/independent variables

Having introduced some background on selecting the best form of the transfer variables, this section turns now to the process of selecting the independent variables (or features, as they are more commonly known in the ML domain) that are inputted into the training algorithm to model/predict the transfer variables. For consistency in terminology, the term features or "(native/augmented) common variables" are used to refer to this component for the remainder of the chapter. The features used in the transfer process must be variables that are common to both the donor and recipient datasets – this means they must be present for all cases in both datasets. These variables may fall into two categories: *native* common variables or *augmented* common variables.

Native common variables exist initially in both datasets and they tend to be socioeconomic and demographic (SED) variables as those are commonly present in most surveys. Often, native common variable categories must be adjusted and recoded across sources to become compatible for use in the transfer process. Augmented common variables are obtained from external active or passive datasets and so must be appended to both the donor and recipient datasets. These external datasets must be able to be joined to the donor and recipient datasets at the individual, household, or geographic level; as such, identifiers at one of these levels must be present across the datasets (statistical matching techniques may be needed at this stage to match identifiers but note that this may increase/compound the error in the subsequent transfer learning process). The number of augmentation datasets used to provide features during the variable transfer process significantly affects the complexity of the transfer process. The remainder of this section details the steps that must be taken to process the datasets from which augmented common variables are obtained.

Firstly, the external datasets from which the features (i.e., feature datasets) are derived may require pre-processing, a critical stage that includes cleaning, validating, and integrating the feature datasets with the recipient and donor datasets. The dataset integration process can be methodologically complex and as noted, may require some use of statistical matching techniques (Shaw et al., under review). Next, the key common variables that will be used from each feature dataset should be examined for missing data, and decisions regarding data imputation or removal may be established. Following this, variables that violate pre-specified correlation (e.g., nearly equal to 1 with other variables) and variance thresholds (e.g., nearly 0) should be removed from the feature sets, as this can

adversely affect the performance of the transfer algorithms. The selection of correlation and variance thresholds may require experimentation and iteration before converging on the best thresholds for a set of features.

Following the pre-processing stage, analysts should then examine the performance of each subset of features and perform dimension reduction as needed. The latter step may only be necessary if the number of features is significantly larger than the number of observations, an occurrence that is often referred to as the curse of dimensionality. In essence, when there are too many features or dimensions, it becomes increasingly difficult to find patterns in the observations as the distances between observations begin to appear equal (Yiu, 2019a). In addition to the curse of dimensionality, the processing time of the transfer process can increase significantly as the number of features increase. Accordingly, dimensionality reduction is used to find the underlying trends or dimensions in the data, and to thereby combine the variables that constitute these trends, hence projecting the data onto a lower dimensional space (Yiu, 2019b). There are many methods and approaches for performing dimension reduction, and analysts may examine multiple approaches to determine the best one for their variable transfer process.

There are further steps that can be taken to expand the feature set and improve the transfer performance. One such step is basis expansion, a process that allows for the augmentation of features with various transformations (e.g., polynomial and interaction terms, natural cubic splines), thereby allowing for non-linearity in the relationship between input and output variables (but expanding the number of features dramatically – for natural cubic splines, the expansion would be by a factor of three for those variables alone, not counting possible interaction terms and other transformations). Keep in mind that after

basis expansion, dimension reduction may again be needed to keep the features at a reasonable number.

Accordingly, it can be seen that as with the transfer variables, there are numerous factors that can be varied during the selection of the features used during the transfer process. Depending on the number of feature datasets used and variations explored (e.g., exploring a range of variance/correlation thresholds or missing data imputation approaches), this component of the transfer process can be the most time consuming and resource intensive, as was the case in the application that is shown in this chapter.

3.4.2.3 Algorithms/functions/models and performance metrics

The third component of the transfer process is the algorithms used to transfer the variables across datasets, and their subsequent performance metrics. For consistency, the term (transfer) algorithms are used to refer to transfer functions/predictive models for the remainder of the chapter.

Transfer algorithms can be either parametric or nonparametric in nature, and as before, analysts may wish to examine several before selecting the choice that yields the highest performance. In this thesis, both types of algorithms are used to provide an example of possible explorations that can be taken. The form of the transfer variable also determines the algorithms that can be tested, with discrete transfer variables requiring classification algorithms or discrete choice models (e.g., ordered logit models) and continuous transfer variables requiring regression algorithms. Further, as expected, the type of (i.e., single versus multi-outcome) algorithms used will depend on whether the transfer variables are being transferred simultaneously or independently. Simultaneous or multi-outcome

algorithms facilitate inter-variable consistency, while independently developed algorithms (i.e., the type used in this chapter) facilitate optimization of performance for each transfer variable.

As will be shown in the application detailed in Section 3.5 of this chapter, there is a focus on supervised ML as the tool of choice for the transfer algorithms (for more information on ML, see: Hastie, Tibshirani, & Friedman, 2016; Bishop, 2006). This is because ML algorithms are able to optimize predictive performance using a large number of inputs or features with an eye toward replicability within other datasets (i.e., through the use of regularization parameters). In addition to these powerful characteristics, many ML algorithms are also nonparametric, meaning that they do not make any assumptions about the input data and their resulting predictions are not constrained by distributional assumptions (Pawlak et al., 2013). Three possible ways of grouping ML algorithms are gradient descent-based algorithms (e.g., elastic net, lasso, ridge regression), distance-based algorithms (e.g., support vector machine (SVM), k-nearest neighbor (kNN), and tree-based algorithms (e.g., random forest (RF), extreme gradient boosting (XGB)). While ML algorithms are considered the best tool of choice for transfer learning, it is possible that traditional regression and choice-based algorithms will perform well-enough depending on the dataset and features; and further, lack of familiarity with ML should not preclude analysts from applying this framework with simpler algorithms.

In addition to exploring various algorithms, parameters within the algorithm development and execution process can also be varied. For example, if using ML algorithms, the hyperparameters that constrain those algorithms can either be set to default values, or alternatively can be tuned using various approaches (e.g., different methods of

cross validation: k-fold, leave-one-out; various methods of parameter testing: random search, discrete grid search, continuous tuning; etc.) to yield potentially better results. Hyperparameters vary across algorithms and are used to control the specific instance of the ML algorithm being applied; for example, this might look like a parameter that penalizes high coefficients in a certain algorithm (as in the case of the regularization parameters in elastic net regression). Meanwhile, in this context, cross validation approaches are used to divide the dataset in various ways so as to test various hyperparameter values on different slices of the data. Lastly, the different approaches for testing combinations of hyperparameters are discussed here. Specifically, random search tuning selects random combinations of hyperparameters to test from the provided values, while grid search tuning performs an exhaustive test of every parameter combination. Continuous tuning tests parameters selected from *ranges* of values provided to the algorithm.

There are also additional feature preprocessing steps that are needed by some ML algorithms. For example, if the input features have vastly different ranges, gradient-descent (e.g., lasso regression, elastic net regression) and distance-based algorithms (e.g., support vector machine and k-nearest neighbor) may require the feature sets to be normalized or standardized. The choice of normalization or standardization depends on assumptions or knowledge regarding the expected distribution of each feature (Bhandari, 2020). In addition, for both ML and traditional modeling algorithms, it is considered best practice to train the algorithms on a portion of the data (training sample/set), and then test them on the remainder (test sample/set). This is known as the training/test split and varying this ratio can influence the final performance. For more information on why training and test sets

should be used during all modeling exercises (i.e., not just for ML algorithms), see (Parady, Ory, & Walker, 2021; Walker, Vij, and Brathwaite, 2019).

Furthermore, depending on the algorithms chosen for comparison, various performance metrics can be used to compare the relative performance of each subset of parameters used. Examples of such metrics include coefficient of determination (parametric models only), mean squared errors (continuous transfer variables only), misclassification error (discrete transfer variables only), and correlations between predicted and observed transfer variable values (a statistical metric). There are a *wide* range of other metrics that can be used in the evaluation of transfer algorithms, and analysts are encouraged to explore these, as various metrics may be better suited to different application contexts (Minaee, 2019).

There is a significant amount of literature in transportation and other disciplines on algorithmic comparisons (e.g., between traditional models and ML algorithms, between various ML algorithms, etc.). The results are generally mixed, with some studies reporting that ML algorithms have superior predictive performance (> 5%), while others find that ML algorithms perform at more-or-less the same (+/- 5% change in predictive accuracy or variance explained) or lower levels (< 5%), relative to traditional modeling approaches. Conceptually, there are numerous reasons why algorithmic comparisons might differ, including differences in: outcomes being predicted, number and types of features/predictors, predictor preprocessing/selection, hyperparameter tuning and selection, cross validation, and differences in comparison metrics used. Support for this can also be found in the literature, with a select group of comparison studies showing that adjusting one or more of the preceding factors can affect the overall final performance of

various algorithms (Feng et al., 2019; Kerckhoffs, Hoek, Portengen, Brunekreef, & Vermeulen, 2019; Wang & Ross, 2018).

Thus, in closing, the process of choosing transfer algorithms and selecting optimal parameters can impact the results observed; however, the test of performance will lie in the usefulness of the transferred variables within the respective field of study. Accordingly, many applications may not need copious experimentation with multiple algorithms and their parameters, as a simple transfer process may suffice.

3.4.3 Integrating across components: transfer variables, features, and algorithms

As one might be able to ascertain at this point, the process of determining the best parameters across all three components of the transfer process is somewhat circular in nature as the analyst has to simultaneously explore parameters that can change within each of the three components of the transfer process in order to select starting points for comparison within each component. Accordingly, based on experience, it is recommended to first conduct some exploratory investigations – initially randomly varying the transfer variables and feature subsets for different algorithms and aiming to get a sense of which algorithm performs best in general at a default level (i.e., keeping all algorithm hyperparameters at the default/recommended values – most programming environments have these values already programmed and/or have documentation about what the values are).

Once a tentatively superior algorithm is selected from this initial exploratory process, we recommend varying the parameters for the transfer variables and determining the form of the transfer variables with the best performance for the chosen algorithm. After

this, the parameters for the features should be varied and the best performing set of features selected. Lastly, the final sets of transfer variables and features can be systematically compared across the chosen algorithms to confirm that the first superior algorithm still stands. At this stage, the analyst may choose to optimize the algorithms, testing a wide range of hyperparameter tuning and cross validation approaches. One might imagine that in the event that another algorithm emerges as better, the analyst could choose to iterate once more through the process, beginning with the transfer variables and moving to the feature sets. The order of approach suggested here is of course flexible and analysts should experiment within the context of their specific applications.

3.5 Transfer-learning application

Having provided an understanding of the parameters that can be varied within each of the three main components of the transfer process, we turn now to an application of the transfer framework.

3.5.1 Overview of application

We apply the method and process discussed in Section 3.4 to develop algorithms (i.e., f_{GDOT} in Equation 1 and Equation 2 at the end of this section) that allow for the transfer of attitudinal variables from the GDOT Survey (donor survey) into the NHTS Georgia sub-sample (recipient survey). Detailed information regarding the NHTS and GDOT datasets can be found in Section 1.3 of this thesis. As a brief refresher, the NHTS dataset contains individual and household- level travel behavior data collected for all 50 states, although the subset used in this thesis is only for the state of Georgia. The NHTS lacks psychometric values that may lend additional insight and improvement in transport forecasting and

planning applications. Meanwhile, the GDOT survey dataset is a research-oriented statewide transportation survey that obtained a wide array of attitudinal data. The goal of the application of this study is to enrich the NHTS dataset with attitudinal data from the GDOT survey.

First, it is necessary to discuss the conditions/assumptions that the framework imposes upon the datasets (see Section 3.4.1). The first major condition encourages spatial and temporal alignment to the greatest extent possible so as to minimize differences in unobserved attributes. The NHTS and GDOT survey were both collected in the state of Georgia in the same general time frame of 2016 to 2018, a fortuitous occurrence that is admittedly difficult to obtain for disparate survey datasets. Next, it is important to get a sense of how similar the datasets are with respect to the distributions of the common variables and/or key sociodemographic attributes. The marginal distributions of the SED common variables that are native to both surveys are shown in Table C1 in Appendix C. Both through visual inspection as well as through significance testing, differences in distributions are found to be small to medium, and as such measures to mitigate these differences are not taken. Furthermore, given that the common variable space is augmented with (literally) thousands of additional common variables, it is plausible to expect that differences between the common variable distributions are controlled for via the large number of variables also present in the transfer algorithms.

As mentioned in the preceding paragraph, both the GDOT and NHTS survey samples used in this application have been augmented with external data sources: (1) targeted marketing (TM) variables that are purchased for all respondents from a commercial data compiler/provider; (2) transit service variables purchased from

AllTransit™; and (3) land use (LU) variables associated with respondents' residential locations and derived from five-year American Community Survey (ACS) estimates and the Environmental Protection Agency Smart Location Database (EPA SLD). Background details on the TM data used in this chapter can be found throughout Chapter 2, and more specifically in Section 2.5.1. The EPA SLD, created in 2013, is a dataset based almost entirely on Census and ACS data that was developed by the EPA Smart Growth program to provide a data resource that could be used to examine location efficiency (Ramsey & Bell, 2014). The ACS, conducted by the U.S. Census Bureau, is a yearly ongoing survey that was designed to augment the Census by providing additional information on smaller samples. The ACS data used in this study represents the five year 2013-2017 ACS estimates (i.e., aggregated across that time; U.S. Census Bureau, 2020). The AllTransit™ data is aggregated from General Transit Feed Specification (GTFS) data by the Center for Neighborhood Technology (CNT), and captures transit connectivity, access, and frequency (Center for Neighborhood Technology, 2019).

To provide context for the application, in Figure 3.4 and Equations 1 and 2, we show application-specific versions of the transfer learning equations and process shown in Figure 3.2. We emphasize that while augmented common variables may improve the transfer performance, it is not necessary to augment the datasets and the transfer-learning process could proceed with just the presence of the native common variables. We present the results of the application from the perspective of both examining the performance of the attitudinal variable transfer process while also providing a detailed sample application and discussion of the transfer learning framework. For the purposes of this chapter, since we are varying and comparing a multitude of parameters across all three components of

the transfer process, we use correlations between the observed and predicted attitudes for the donor dataset as the only performance metric as it is easily comprehensible which aids in simplifying the many parameter comparisons that are shown. As before mentioned, we recommend that researchers examine all applicable metrics, and select the one that best summarizes the intended information and trends that are pertinent to the research question at hand.

$$Attitudes_{GDOT} = f_{GDOT}(CV_{GDOT}, augCV_{GDOT}) + \varepsilon_{GDOT} \quad (1)$$

$$\widehat{Attitudes}_{NHTS} = f_{GDOT}(CV_{NHTS}, augCV_{NHTS}) \quad (2)$$

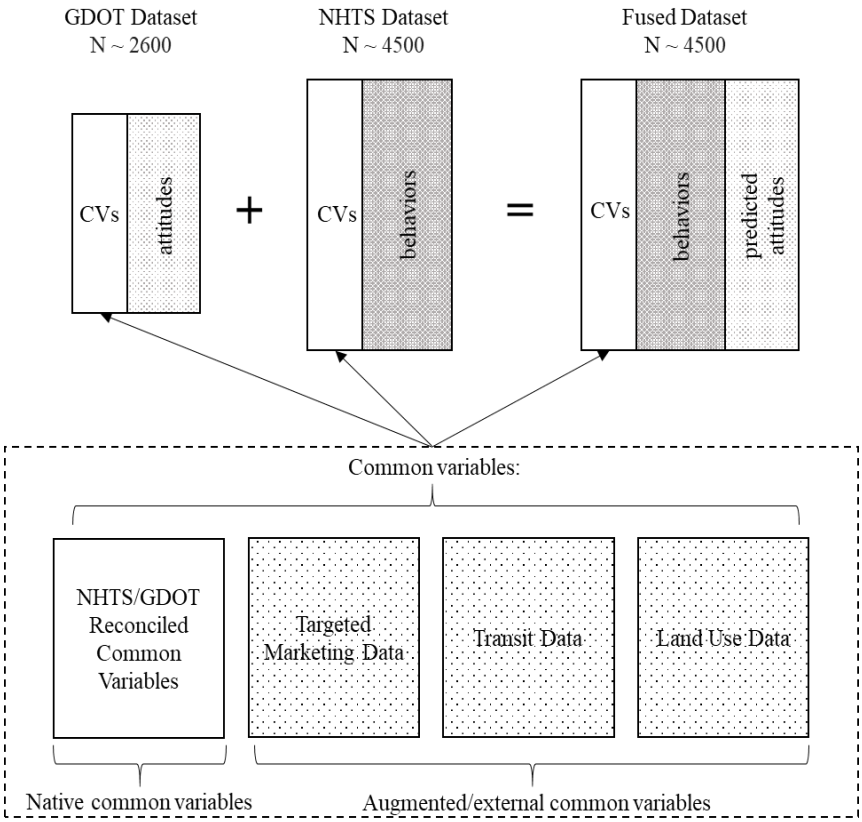

Figure 3.4. Application-specific transfer learning framework

3.5.2 Attitudinal transfer variables

In this application, the original transfer variables consist of 36 discrete attitudinal indicator variables with a five-point ordinal rating scale ranging from "Strongly disagree" to "Strongly agree" (See Table C2 in Appendix C for attitudinal indicators/statements in this analysis). Informed by initial explorations, we choose to transfer latent attitudinal

constructs which are continuous variables developed using these discrete indicator variables (Malokin, 2019). **Figure 3.5** illustrates a latent attitudinal construct used in this study and its indicator statements. These latent attitudinal constructs (i.e., now our transfer variables) are transferred independently of each other, as is the recommended approach when aiming to obtain the best possible result for each transfer variable. Since the methodology for developing latent constructs is outside the scope of this chapter, interested readers are directed to Stevens (2009) for further explanation.

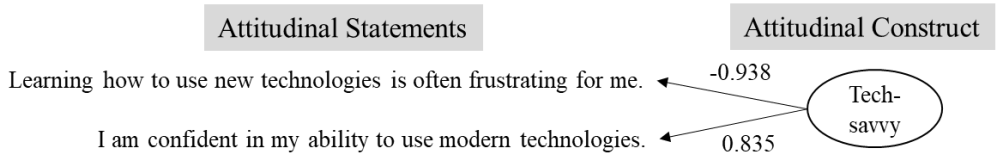

Figure 3.5. Example of attitudinal indicator statements and corresponding latent construct

Four different forms of transfer variables are extracted for comparison. First, we use two different latent structure identification methods: EFA and CFA (Table C2 in Appendix C shows both EFA and CFA results). Next, we extract both six and fifteen-factor solutions for each method (note that factor and latent construct are interchangeable terms in this context, and we use them as such). The six-factor solution was developed based on initially applying the eigenvalue greater than 1 rule to the extraction sums of squared loadings, while the fifteen-factor solution was developed based on applying this same rule to the initial eigenvalues. These rules represented starting points for finding the final solution, and thereafter interpretability, simple structure, communalities, and factor correlations were used to develop and tune the final solutions (Stevens, 2009). There are numerous approaches to determining the number of factors to extract, and in the case of

this study, – our primary goal was to obtain an interpretable solution that had more general constructs (i.e., the six-factor solution) and an interpretable solution with more specific constructs (i.e., the fifteen-factor solution).

Figure 3.6 provides a graphical comparison of differences in transfer results obtained across the four transfer variable solutions tested. As is conventional, the factor names shown on each graph are developed based on the indicators/statements that have the highest loadings for these factors (see Table C2 in Appendix C). Based on initial explorations and iterations (see Section 3.4.3), the algorithm used in this section is elastic net regression with discrete hyperparameters (penalty terms: alpha and lambda) tuned using grid search. The features used are all of the native common variables, alongside all TM variables, and the subset of land use variables found to be the overall best choice in Section 3.5.3.2.2 (i.e., the set of variables that accounts for 50% of the variance in the EPA and All Transit dataset and 50% of the variance in the ACS dataset). These choices were made during the second round of iteration, once the overall best algorithm and feature subsets had been determined. Overall, we see only modest differences in performance between EFA and CFA, as well as between the six and fifteen-factor solutions, although as expected, this varies by transfer variable. Due to these modest differences, moving forward, the **fifteen-factor transfer variable EFA solution** is used when exploring other components and parameters in the transfer process.

Lastly, as shown on the graphs here and throughout the remainder of the results section (Section 3.5), the attitudinal constructs have been subjectively categorized into four domains so as to explore any subject-level trends that might be occurring in performance differences (and thus which may require further investigation). The results are ordered from

highest to lowest performance based on the transfer variables within each respective domain. In general, across all solutions shown, lifestyle-oriented attitudinal variables tend to have the highest performance, followed by travel, land use, and finally, personality-oriented variables. In the fifteen-factor solution, we see that correlations between the predicted and observed values for the work-oriented construct is 0.51 for the EFA solution, which signifies an R^2 value of 0.26. This means that with the features used in this prediction, we are able to explain 26% of the variance in this attitudinal construct; this is considered to be a very good R^2 value for psychometric variables of this type.

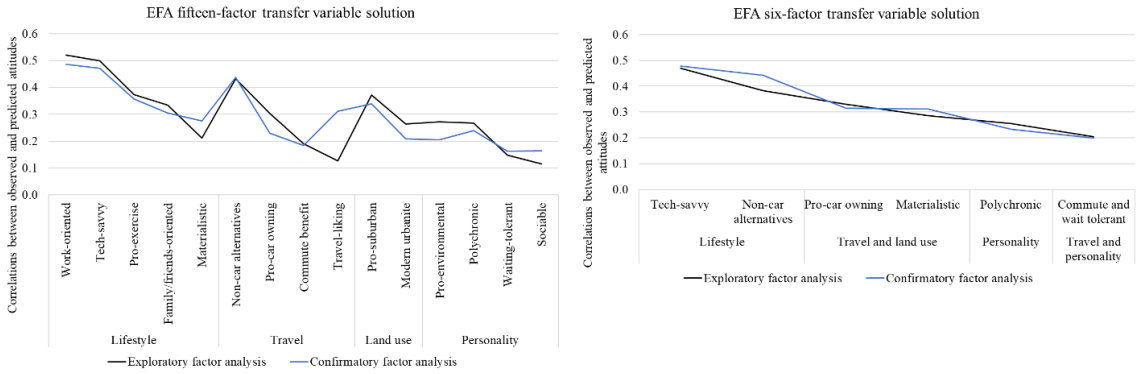

Figure 3.6. Comparison of EFA and CFA results for fifteen and six-factor attitudinal constructs

3.5.3 Features/inputs for attitudinal variable transfer

We now discuss the process used to obtain the features that serve as inputs to the transfer algorithms. As discussed, (see Figure 3.4 in Section 3.5.1) the features used in this application are composed of both native (to the donor and recipient datasets) and augmented common variables. In Figure 3.7, we summarize the steps that were taken to develop a final set of features for each common variable set for the variable transfer process presented in this chapter. Note that the optimal pathway through the process for each

transfer variable varies, but in this thesis, for ease of comprehension, we choose the overall best path considering all the transfer variables. Nonetheless, where appropriate, throughout this section, we strive to illustrate the difference in results that can occur while using various subsets for different transfer variables. We also select the optimal performance for each transfer variable (without holding the subsets constant) to use for comparison purposes in the discussion (Section 3.6).

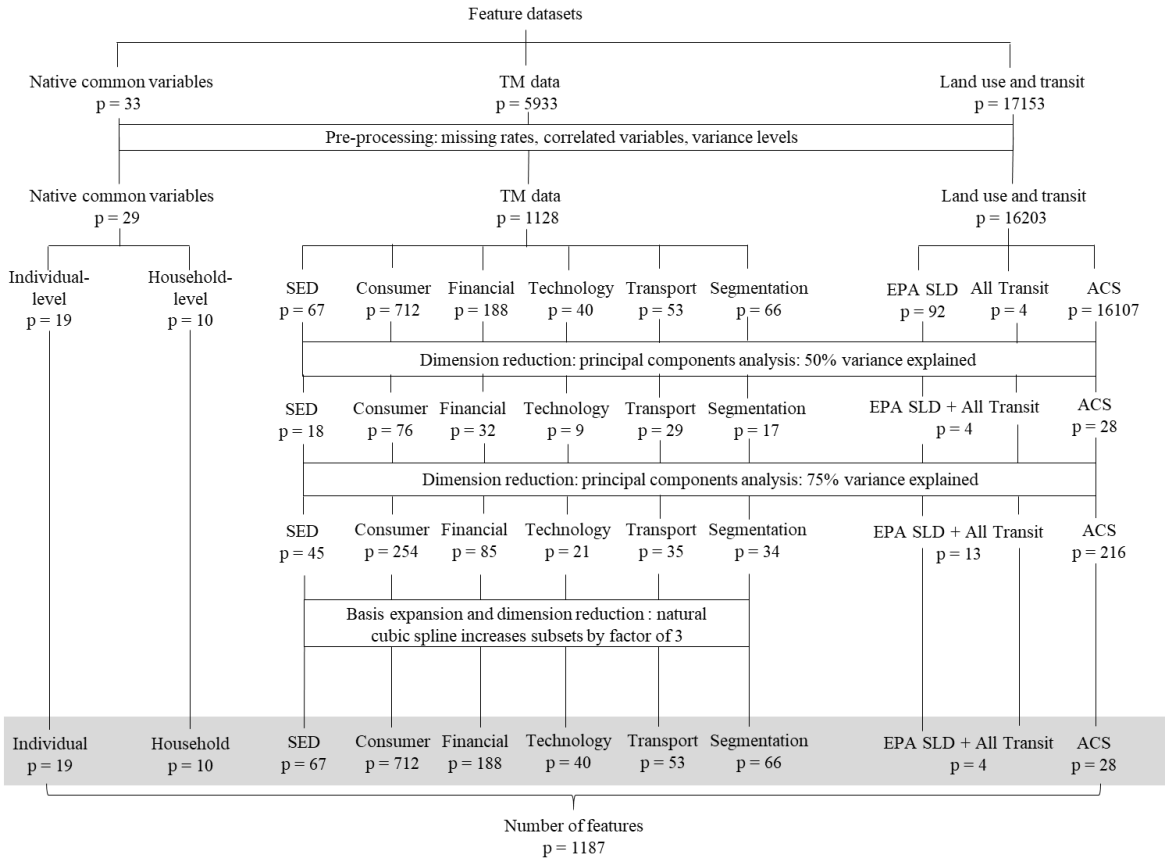

Figure 3.7. Sample process for determining the final sets of features for use in transfer process

(grey represents the optimal feature set selected for use in the application detailed in this chapter)

3.5.3.1 Native common variables

As mentioned, native common variables tend to be SED variables. In Table C4 in Appendix C, a list of the native common variables used in this study is provided as well as the recoding that was necessary across the donor and recipient datasets to harmonize the common variables. In the results shown here, these variables were separated out into individual and household-level variables to provide some insight into how various types of features can result in performance differences. The individual-level variables include core SED variables like gender, age, education, race, worker status, driver status, and medical conditions. The household-level variables include attributes that are typically reported at the household level: household income, number of drivers in household, and number of household individuals in various age groups. As shown in Figure 3.8, individual-level variables perform better than the household-level variables for most of the attitudinal variables being transferred. This is significant because it means that even in the absence of variables like income, the attitudinal variables of interest in this study (with the intuitive exception of materialistic, which is an attitude that is correlated with household income) can be transferred almost as well as if household-level variables were not available in the common variable set. In the case of this application, given that the performance is best when using both household and individual-level variable subsets together, we move forward retaining all the native common variables tested (i.e., the overall set).

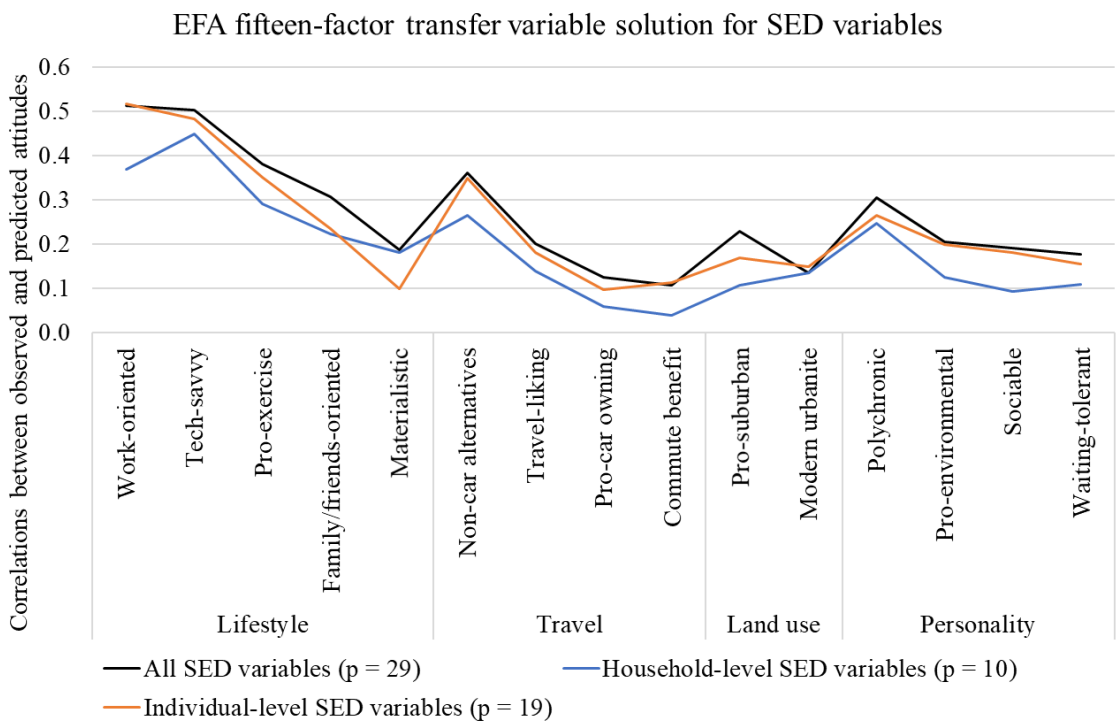

Figure 3.8. Transfer learning results when using native common variables

3.5.3.2 Augmented common variables

As noted, both the donor and recipient datasets were augmented with external features from four additional data sources. This was done to illustrate the potential of using external data sources to support the variable transfer process in the cases where the transfer variables themselves are not available for direct appending to the recipient dataset. Recall that one form of survey data enrichment is to directly augment the data with variables from passive, big data sources (i.e., using identifiers for exact record linkage). However, many types of variables (like psychometric variables) are not available from these passive data sources, making it necessary to obtain them from other surveys and thereby motivating the process shown here.

3.5.3.2.1 Targeted marketing data

The TM dataset used in this study consists of 5933 variables that we classified into the following categories: SED, consumer, financial, transport, technology, and segmentation variables (see Section 2.5.1 for more details). We preprocessed the variables, checking for and removing variables with high rates of missingness and low variation. Next, we removed variables with high rates of correlation with other variables in the respective subset. After the pre-processing steps were applied, the total number of TM variables remaining is 1128 variables.

As shown in Figure 3.7, the next step when dealing with a large number of features (as is typical when using big data) is dimension reduction. While there are many dimension reduction approaches, in this chapter we applied PCA which is a linear approach that uses singular variance decomposition to project the data to a lower dimensional space. In Figure 3.9, we show differences in results obtained when using the entire subset of TM variables ($p = 1128$), as well as dimension reduced subsets of TM components that explain 25% ($p = 6$), 50% ($p = 100$), and 75% ($p = 343$) of the variance in the data, respectively. In general, the subset of components that explains 50% of the variance ($p=100$) outperforms the other dimension reduced subsets, with this subset performing almost as well as when using all TM variables. This may be because the number of variables in this subset falls into a "sweet spot" with regards to number of variables relative to information provided (see the curse of dimensionality concept discussed in Section 3.4.2.2, a phenomenon that explains why the use of too many features, relative to the number of observations, can sometimes result in decreased performance). Based on these results, moving forward in the parameter testing

and iteration process we test the full TM variable subset, as well as the subset of TM variables that explains 50% of the variance in the original set of TM variables.

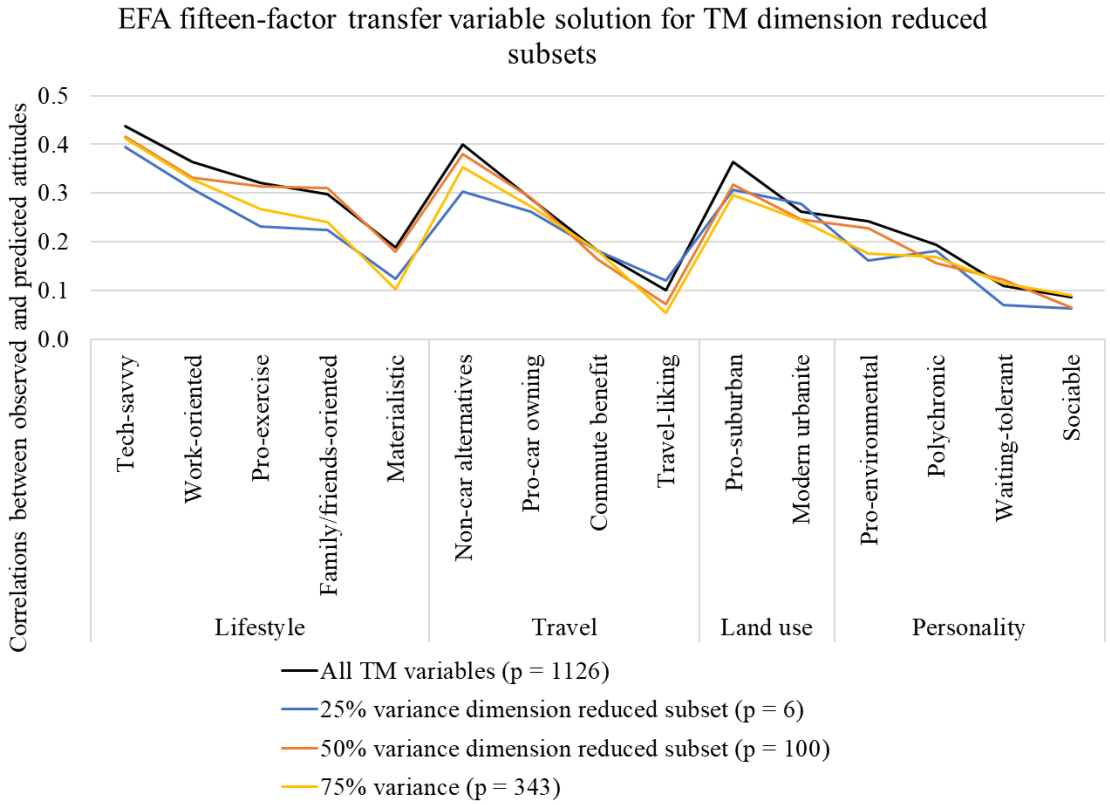

Figure 3.9. Transfer learning results when using various subsets of TM variable principal components

Next, in Figure 3.10 we show the transfer performance of various categories of TM variables (see Table B1 in Appendix B.1), which again allows us to understand how varying *domains* of common variables can contribute differently to the transfer process. We see that TM *consumer* variables outperform the other categories of TM variables for eight of the 15 transfer variables. The consumer subset of variables represents the largest category of TM variables, which may be one potential explanation why this category performs better relative to the other TM variable categories (i.e., more information). The TM *segmentation* variables outperform the other categories of TM variables for four of the

15 transfer variables (i.e., tech-savvy, polychronic, travel-liking, and materialistic variables), which is again intuitive given that these variables already represent membership in clusters that are formed by intelligently identifying population groups having similar profiles on combinations of multiple lifestyle-related variables. Meanwhile, the TM *transport* subset performs best for the sociable construct, while the TM *SED* subset perform best for the waiting-tolerant construct, and lastly, the TM *financial* subset performs best for the modern urbanite construct.

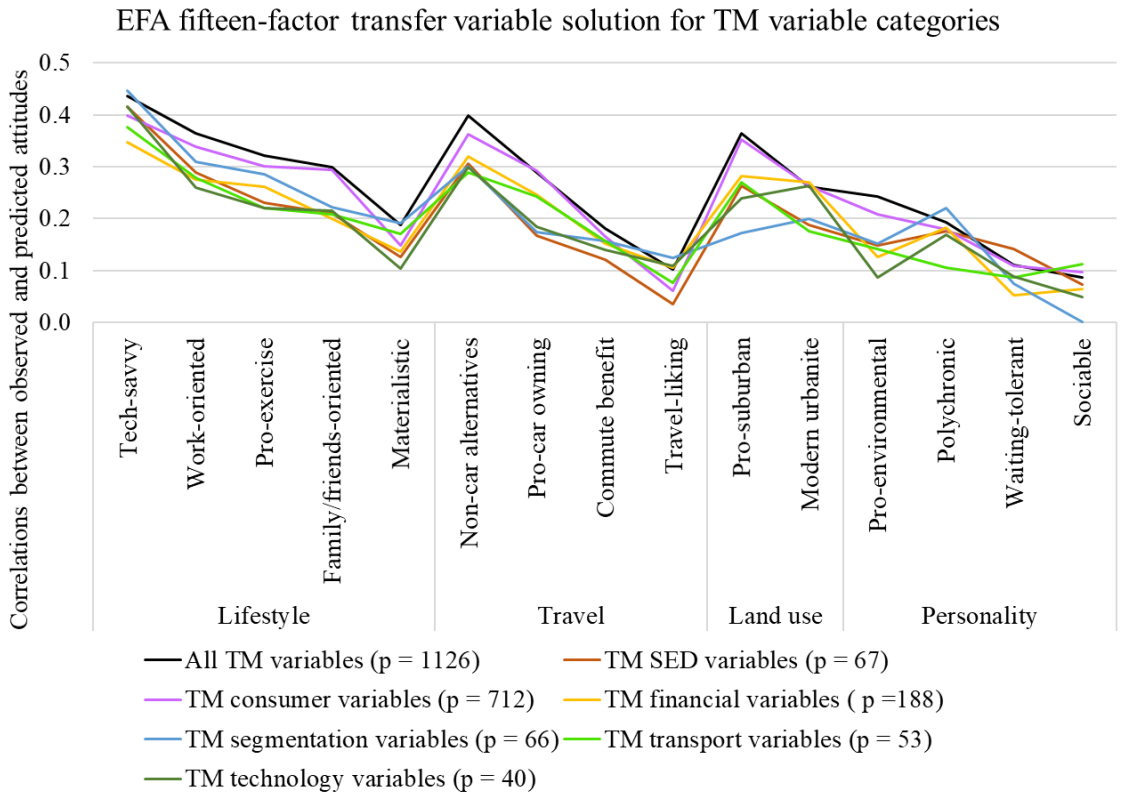

Figure 3.10. Transfer learning results when using various categories of TM variables

Finally, in Figure 3.11 we show the relative performance differences after applying basis expansion through the use of natural cubic splines to the dimension reduced variable subsets that explain 25% and 50% of the variance present in the TM data (see Figure 3.9),

respectively. We see that there is variation across the transfer variables, suggesting that some transfer variables are more likely than others to occur as a result of nonlinear relationships with some TM variables. For example, the materialistic, pro-car owning, travel-liking, pro-suburban, polychronic, and sociable transfer variables appear to have slightly improved performance when creating natural cubic splines from the full (standardized) TM dataset, although in most cases the improvement is indeed very small.

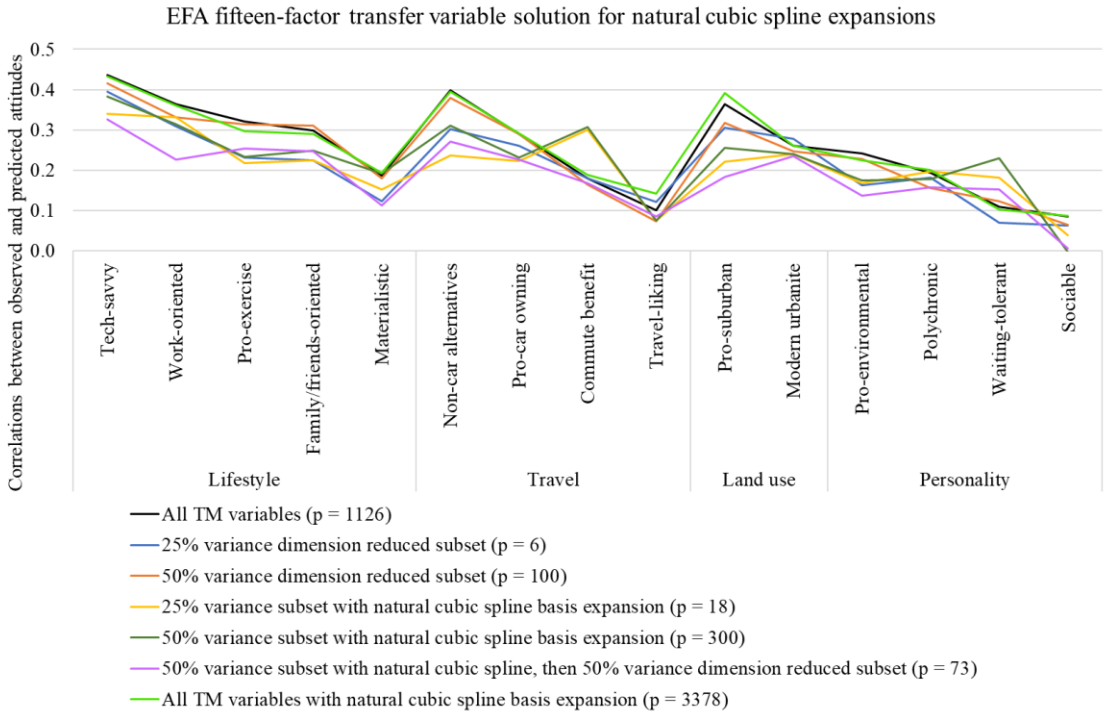

Figure 3.11. Transfer learning results with basis expansion for TM variables and TM subsets

In closing, in Table 3.1 we provide a summary of the subset of TM variables that provided the best observed results for each transfer variable. We see that the full set of TM variables, as well as the set of TM variables with natural cubic spline expansions performed best for 11 of the 15 attitudinal transfer variables. Given that the differences between the set of all TM variables with natural cubic spline expansions and the full set of TM variables

are quite small (see Figure 3.11), moving forward, **the overall best subset of TM variables to use would be the full set of TM variables or the dimension reduced set that explains 50% of the variance as this was shown to perform relatively close to the full set (Figure 3.9)**. Not only do these latter two datasets perform almost as well as the natural cubic spline expansion set, but they are significantly smaller, meaning that computational time for algorithm development and execution is significantly reduced. This demonstrates that while performance is one metric to keep in mind, it is not the only measure that analysts might use to select the best feature set.

Table 3.1. Summary of best performing TM data subsets for attitudinal transfer variables

Domain	Attitudinal construct	Correlation between observed and predicted attitudes	TM data subset	No. of variables in subset (p)
Lifestyle	Tech-savvy	0.446	TM segmentation variables	66
	Work-oriented	0.365	All TM variables	1126
	Pro-exercise	0.322	All TM variables	1126
	Family/friends-oriented	0.310	50% variance dimension reduced TM subset	100
	Materialistic	0.194	All TM variables with natural cubic spline basis expansion	3378
Travel	Non-car alternatives	0.399	All TM variables	1126
	Pro-car owning	0.292	All TM variables with natural cubic spline basis expansion	3378
	Commute benefit	0.306	All TM variables with natural cubic spline basis expansion	3378
	Travel-liking	0.143	All TM variables with natural cubic spline basis expansion	3378
Land use	Pro-suburban	0.392	All TM variables with natural cubic spline basis expansion	3378
	Modern urbanite	0.278	25% variance dimension reduced subset	6
Personality	Pro-environmental	0.242	All TM variables	1126
	Polychronic	0.220	TM segmentation variables	66
	Waiting-tolerant	0.231	All TM variables with natural cubic spline basis expansion	3378
	Sociable	0.112	TM transport variables	53

3.5.3.2.2 Land use and transit data

The land use (EPA SLD and ACS data) and transit data were processed together because they are similar in nature and combining datasets aids in simplifying the number of parameter comparisons necessary. Further, due to high rates of missingness in the transit data, only four variables were retained; these represent primarily transit performance indices and so were available across all cases. As before, processing steps for the land use and transit data are summarized in Figure 3.7 where we see that dimension reduction was only deemed necessary for the ACS data due to the large number of variables present. We note that all variables in this subset of features are available at geographic levels (rather than individual or household levels as with the prior features discussed). Specifically, the EPA SLD, Transit, and most ACS variables are available at the block group level, while other ACS variables are available only at the census tract level (i.e., larger, less precise area relative to block group; may be deemed necessary for more sensitive variables). Accordingly, when appending land use variables, it is necessary to have the residential location of each record at least up to a certain classification; this typically requires the analyst to geocode the provided residential locations using a geocoding service such as Google application programming interfaces (API).

Given that the land use and transit data together comprised three datasets, at this point it may be increasingly apparent that there are many combinations of data subsets that could be tested here. For example, one could apply dimension reduction techniques to all three of the EPA SLD, All Transit, and ACS datasets, and test subsets with varying numbers of components extracted across datasets. At the simplest level, one could test each of the three subsets that comprise this group of external variables individually, varying

the number of dimensions tested for each subset – but note that even this exploration would yield between six to nine different datasets. Accordingly, as in prior sections, for comprehension, we could not show all possible combinations tested. Instead, we seek only to give examples of what could be varied (for analysts who may wish to replicate this method), as well as to summarize the best results.

First, Figure 3.12 contains a summary of results across the EPA SLD and All Transit datasets, with the results showing that there is low variation in performance across the feature subsets examined. Overall, it can be said that the EPA SLD and All Transit combined subsets show better performance results across most of the transfer variables relative to EPA SLD alone. Among the EPA SLD and All Transit subsets, the full set of variables and the dimension reduced variable subset that explains 75% of the variance perform similarly and have the highest performance.

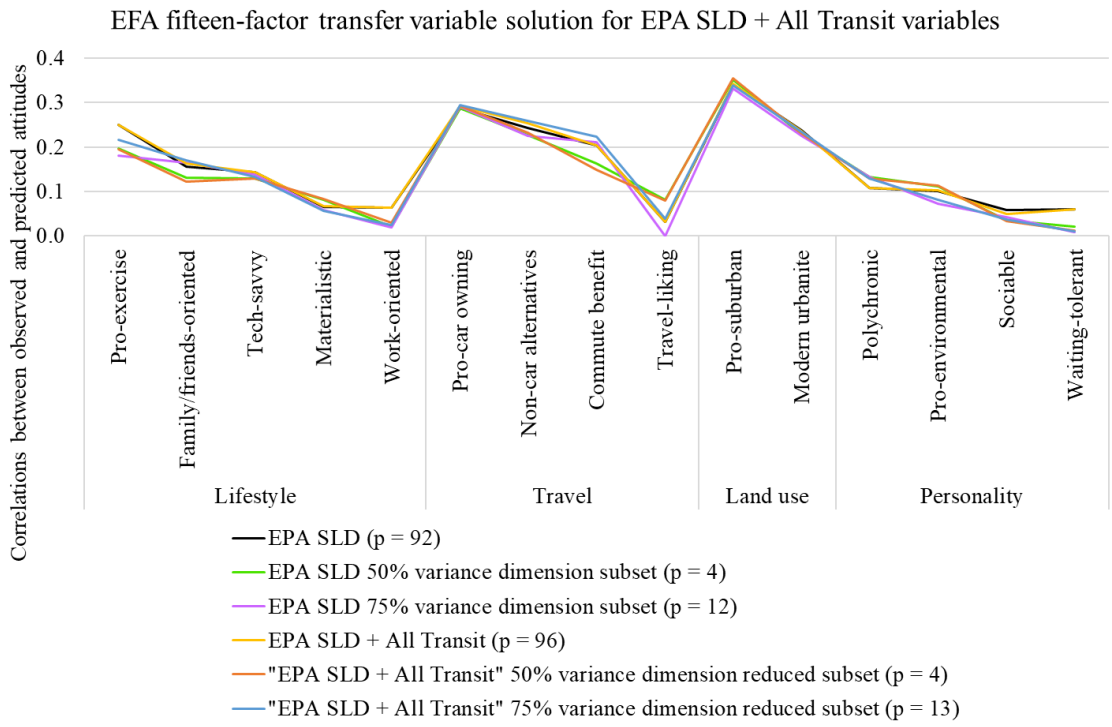

Figure 3.12. Transfer learning results when using EPA SLD and All Transit variables

Next, Figure 3.13 summarizes the performance observed for various subsets of ACS variables, and also provides a comparison of these results relative to the EPA SLD and All Transit dataset. We see that at least one of the ACS datasets outperforms the EPA SLD and All Transit datasets across nine of the transfer variables with the exceptions being non-car alternatives, commute benefit, pro-car owning, pro-suburban, modern urbanite, and waiting-tolerant attitudes. Overall, the full ACS dataset outperforms the ACS dimension reduced subsets, while the subset that explains 50% of the variance outperforms the subset that explains 75% of the variance, findings that mirror those found in the TM feature exploration.

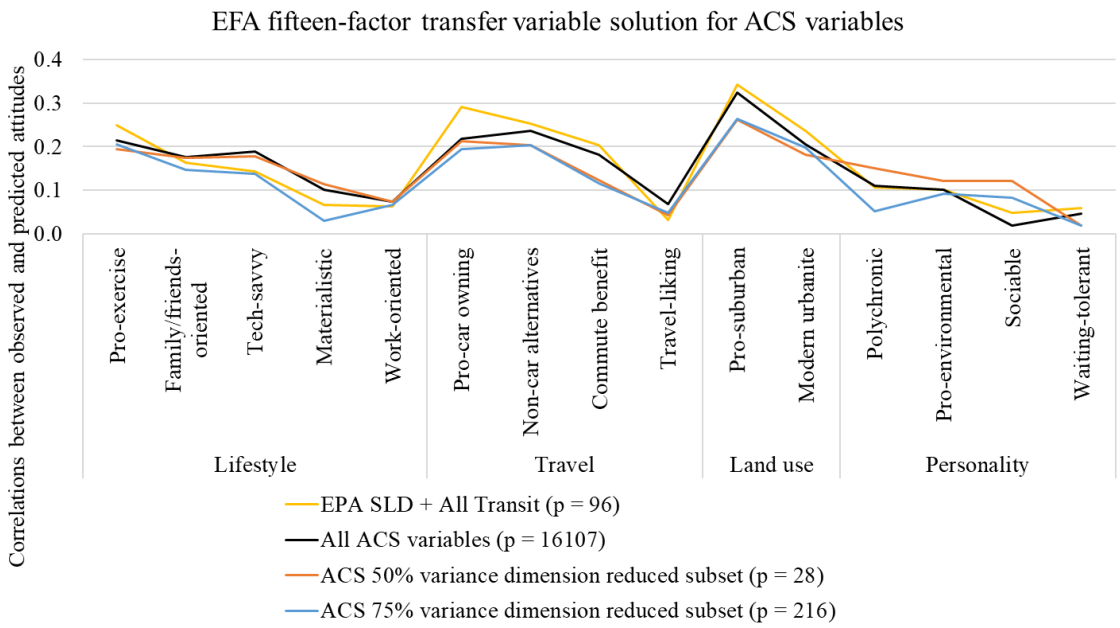

Figure 3.13. Transfer learning results when using ACS, EPA SLD, and All Transit variables

Lastly, in Figure 3.14 we show overall results from integrating across the three feature sets used in this section (EPA SLD, All Transit, and ACS). Although it may be difficult to see from the chart due to the relatively small differences (see Table 3.2 for values), it is the dimension reduced datasets with combinations of features that account for the best results for eight of the transfer variables, with the "EPA and All Transit 50% variance dimension reduced subset and ACS 50% variance dimension reduced subset" accounting for the highest performance for four transfer variables. This may be due to a positive tradeoff between the amount of information provided (i.e., variance explained in original data) and the number of variables, which is relatively small (p = 32) relative to the full combined land use dataset (p = 16203). Due to the high performance of this subset, this is the land use subset that is used when performing additional parameter comparisons.

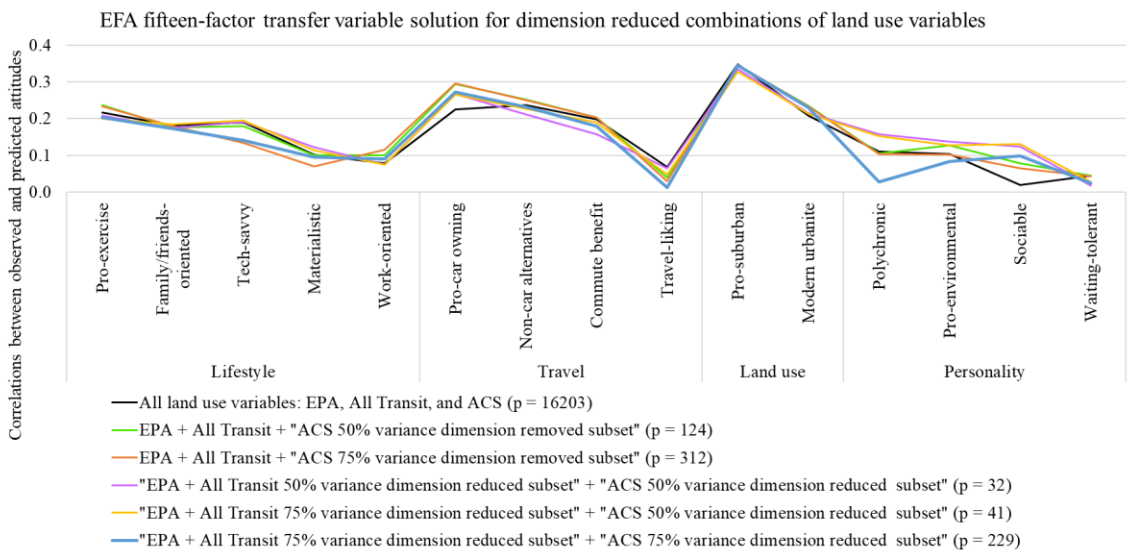

Figure 3.14. Transfer learning results when integrating across land use datasets

Table 3.2 shows that for the land use dataset, seven different subsets performed best for various transfer variables; thereby making it difficult to select one subset as the overall best performer. Along with the earlier results shown for the TM data, this serves to illustrate the importance of optimizing for each transfer variable individually, and also demonstrates the magnitude of the effort that can be expended should the analyst wish to extract the best possible results for each variable being transferred.

Table 3.2. Summary of best performing land use data subsets for attitudinal transfer variables

(grey highlighted cell indicates the best performance observed amongst the compared subsets)

Domains	Attitudinal constructs	EPA SLD	EPA SLD 50% variance dimension subset	"EPA SLD + All Transit" 50% variance dimension reduced subset	EPA SLD + All Transit 75% variance dimension reduced subset	EPA + All Transit + "ACS 75% variance dimension removed subset"	"EPA + All Transit 50% variance dimension reduced subset" + "ACS 50%	"EPA + All Transit 75% variance dimension reduced subset" + "ACS 50%
Lifestyle	Pro-exercise	0.250	0.197	0.194	0.216	0.233	0.209	0.200
	Family/friends-oriented	0.155	0.132	0.122	0.171	0.180	0.174	0.185
	Tech-savvy	0.143	0.129	0.128	0.133	0.134	0.190	0.195
	Materialistic	0.065	0.081	0.083	0.056	0.070	0.122	0.115
	Work-oriented	0.063	0.021	0.030	0.025	0.115	0.075	0.075
Travel	Pro-car owning	0.287	0.287	0.293	0.295	0.296	0.266	0.266
	Non-car alternatives	0.242	0.227	0.232	0.258	0.250	0.212	0.227
	Commute benefit	0.204	0.163	0.149	0.224	0.203	0.157	0.190
	Travel-liking	0.032	0.081	0.080	0.039	0.030	0.065	0.048
Land use	Pro-suburban	0.341	0.351	0.354	0.339	0.347	0.334	0.328
	Modern urbanite	0.237	0.229	0.230	0.234	0.233	0.214	0.213
Personality	Polychronic	0.107	0.133	0.129	0.128	0.104	0.158	0.152
	Pro-environmental	0.101	0.111	0.114	0.082	0.101	0.137	0.127
	Sociable	0.059	0.035	0.034	0.036	0.064	0.124	0.130
	Waiting-tolerant	0.059	0.021	0.011	0.011	0.042	0.017	0.027

In closing, Figure 3.15 illustrates the best performing results (the cells highlighted in grey in Table 3.2 correspond to the black line in Figure 3.15) relative to the original land use datasets examined. This figure also emphasizes that the performance differences amongst different subsets of land use variables are much smaller than in the other feature sets examined (e.g., the native common variables and TM variables). **To reiterate, moving forward the subset of land use variables that will be used in further explorations will**

be the set of variables that accounts for 50% of the variance in the EPA and All Transit dataset and 50% of the variance in the ACS dataset.

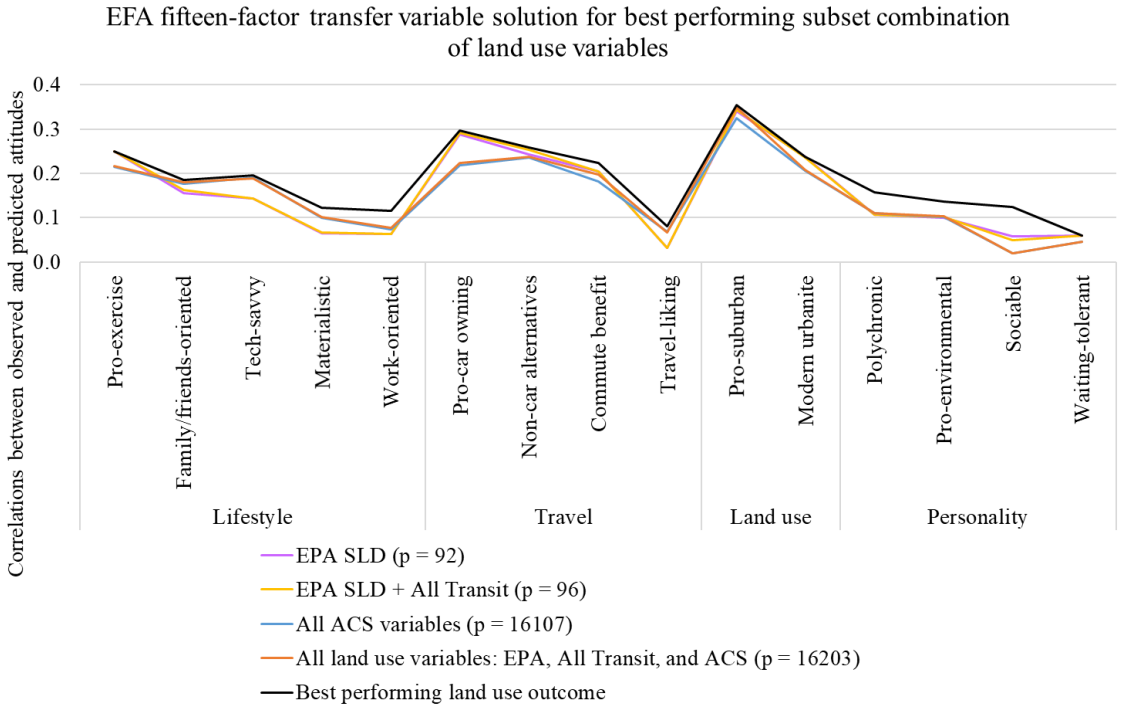

Figure 3.15. Transfer learning results for best performing land use outcome

3.5.4 Algorithms for attitudinal variable transfer

Having discussed the transfer variables and features, the algorithms or models that link these variables are now discussed. Thus far, the results shown in Sections 3.5.2 and 3.5.3 have used the *overall* best performing algorithm – elastic net regression – that was identified as a result of the process that shown in this section. "Overall best performing" means elastic net regression performed best for more of the transfer variables than any other algorithm to which it was compared. As mentioned in Section 3.4.3, the process of determining the best parameters is somewhat circular. For example, without having explored the algorithms that are optimal for the application in this chapter, it would not

have been possible to provide a digestible comparison of the parameters that can be varied in the prior two sections (as we would have to provide the performance results for all combinations of those parameters across all algorithms being explored). Resultingly, the process suggested in Section 3.4.3 was followed, first varying the features and transfer variables across different algorithms to select the algorithm that appeared to be the best performing, in this case, the elastic net algorithm. Having now finalized the best performing transfer variables and feature subsets, we provide a comparison of algorithmic performance across these subsets.

For this application, a range of ML algorithms as well as basic linear regression models are tested (see Figure 3.19). The focus is on gradient descent-based and tree-based algorithms rather than distance-based algorithms in this thesis as they are typically more powerful. As with the other components, there are many parameters that can be varied that may affect the final transfer performance once the algorithms are selected. With regards to input processing prior to algorithm development, in the case of the data used in this application, it was found that some transfer attitudes performed better with normalization versus standardization (and vice versa for other attitudes). Specifically, in this case, normalization was done (feature by feature) by subtracting the minimum value across all cases of a given feature from the feature value for the case in question, divided by the difference between the maximum and minimum values of that feature across all cases (yielding features that have been rescaled to have values between 0 and 1). Because normalization does not make assumptions about variable distribution and also shows better performance for some of the attitudinal transfer variables tested, this approach was used

for processing the training and test sets within the algorithmic training process across all algorithms. There is certainly room for additional exploration in this regard.

Most importantly, there are numerous approaches for tuning the hyperparameters of various ML algorithms that could yield performance differences; as before noted, some of these include: random search tuning, grid search tuning, and continuous (i.e., non-discrete) tuning. In this study, grid search hyperparameter tuning is performed for the ML algorithms using 10-fold cross validation (CV), as explorations showed *overall* superior transfer performance for this dataset. However, it is also noted that continuous and random search tuning approaches perform better for some of the attitudinal transfer variables (see Figure 3.16).

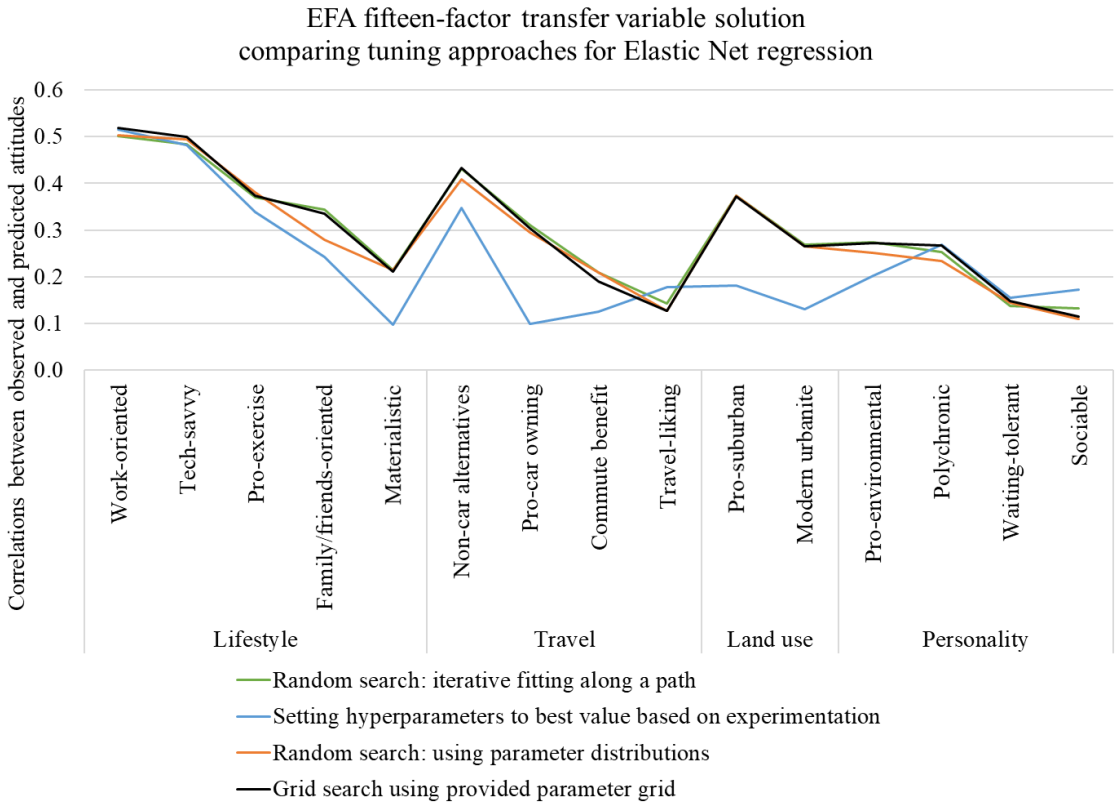

Figure 3.16. Comparison of performance across various hyperparameter tuning approaches

In addition, training/test set split ratios and random variations in how the dataset is divided to yield the training/test sets could also affect the outcomes observed. These parameters would affect the results of transfer learning even in the presence of non-ML transfer algorithms, and we recommend that all analysts explore the performance effects of varying these parameters. In the application shown here, the algorithms were trained on 80% of the data and tested on the remaining 20% of the data (80/20 split) as is the convention in the ML domain (The Data Detective, 2020). However, it is difficult to have a general rule as to the best split since this is very dataset dependent; those interested are referred to Hastie et al. (2016). As shown in Figure 3.17, training and testing the algorithm on all of the data (i.e., the same subset) produces better results because the algorithm is being optimized or overfit for the specific dataset being used. Using separate training and test sets allows analysts to explore how well algorithms are performing on data they have not yet seen. However, note that the test set is still often from the same population or data source as the training set (as is the case in this paper), and so may yet yield better results that what one would see if the algorithm were tested on an entirely new set of data.

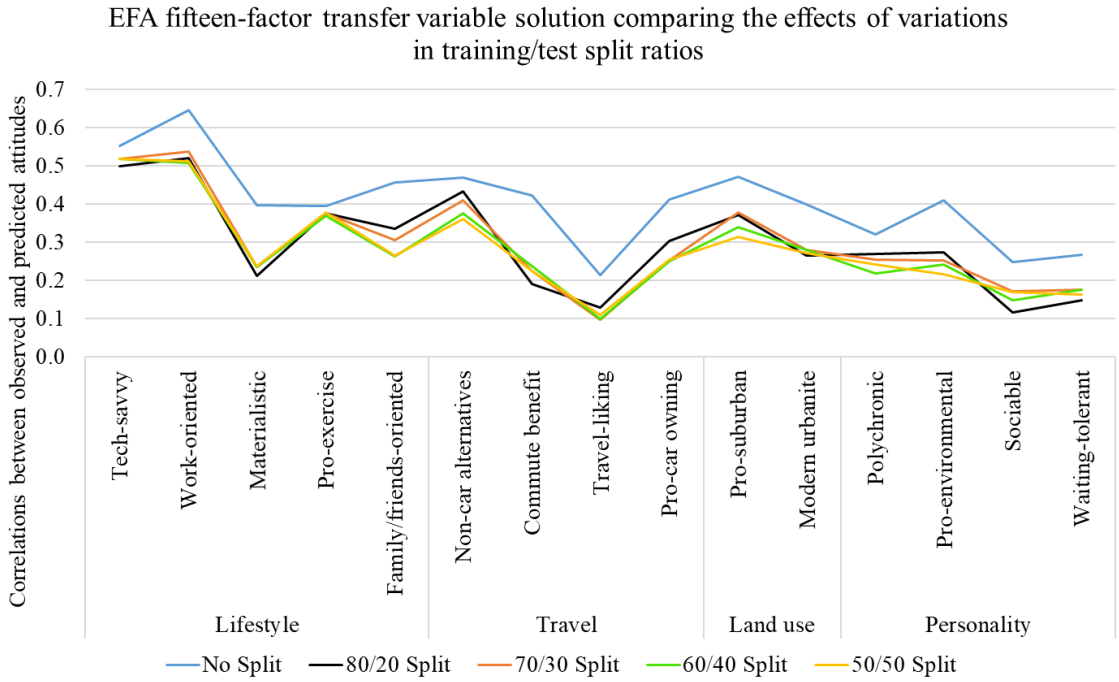

Figure 3.17. Comparison of performance across various training/test set split ratios

Figure 3.18 shows a few examples of the variation in results that can occur due to random differences in how the dataset is split for training and testing. Overall, the general trends across transfer variables remain constant, which supports the stability of the results shown in this chapter. One approach that researchers have used is to average performance metrics across several random splits to yield more generalized results. However, note that the final imputed values for the test dataset in this case come from only one split – and unless the downstream analysis facilitates the use of multiple imputation values, it would be unwieldy to have several outcomes. In this study, the random split was controlled across algorithms in order to ensure that the results shown are comparable across different algorithms, feature sets, and transfer variables. In addition, given the wide array of parameters varied as well as the fact that the resulting predicted values are used as inputs for a range of methods during internal and external validation, it would be all but

impossible to use multiple imputation. Nonetheless, this is an area of investigation that should certainly be explored further in the application of this method.

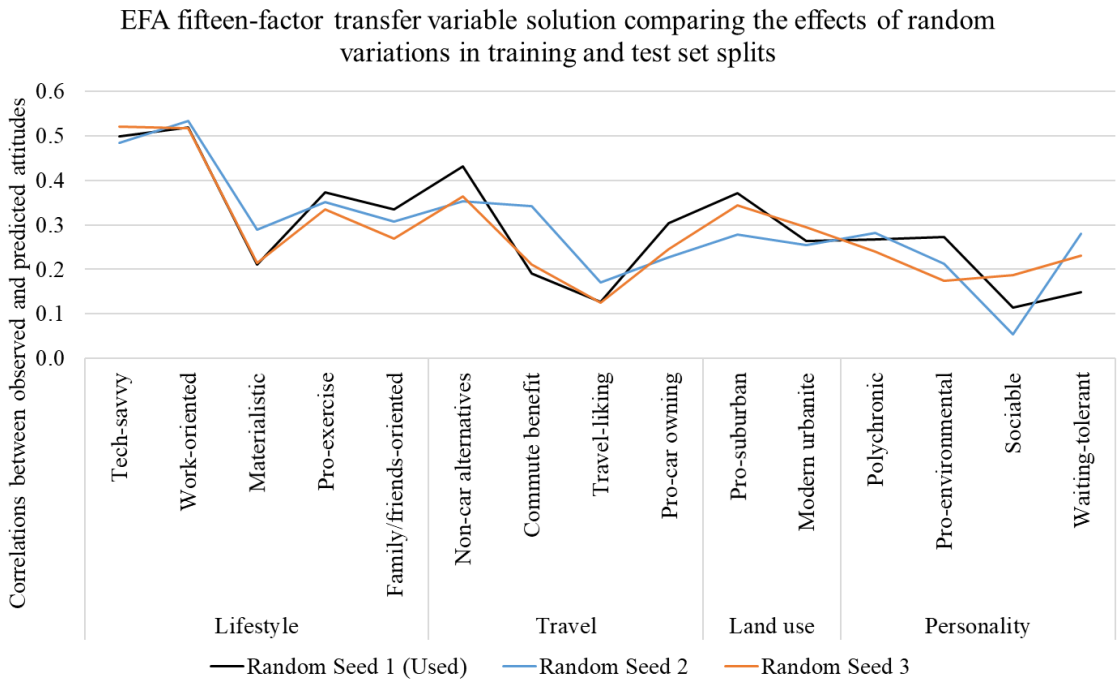

Figure 3.18. Comparison of performance across random variations in training and test set splits

Figure 3.19 illustrates the transfer learning results across the algorithms tested for this chapter. We optimized the tuning approaches by algorithm but kept constant all other parameters across algorithms (with the exception being linear regression which did not have parameters to be tuned). The features are normalized for all algorithms, and the training/test split is 80/20. It is seen that elastic net regression continues to outperform the other algorithms for a majority of the variables, with extreme gradient boosting and support vector regression doing better for a handful of other variables.

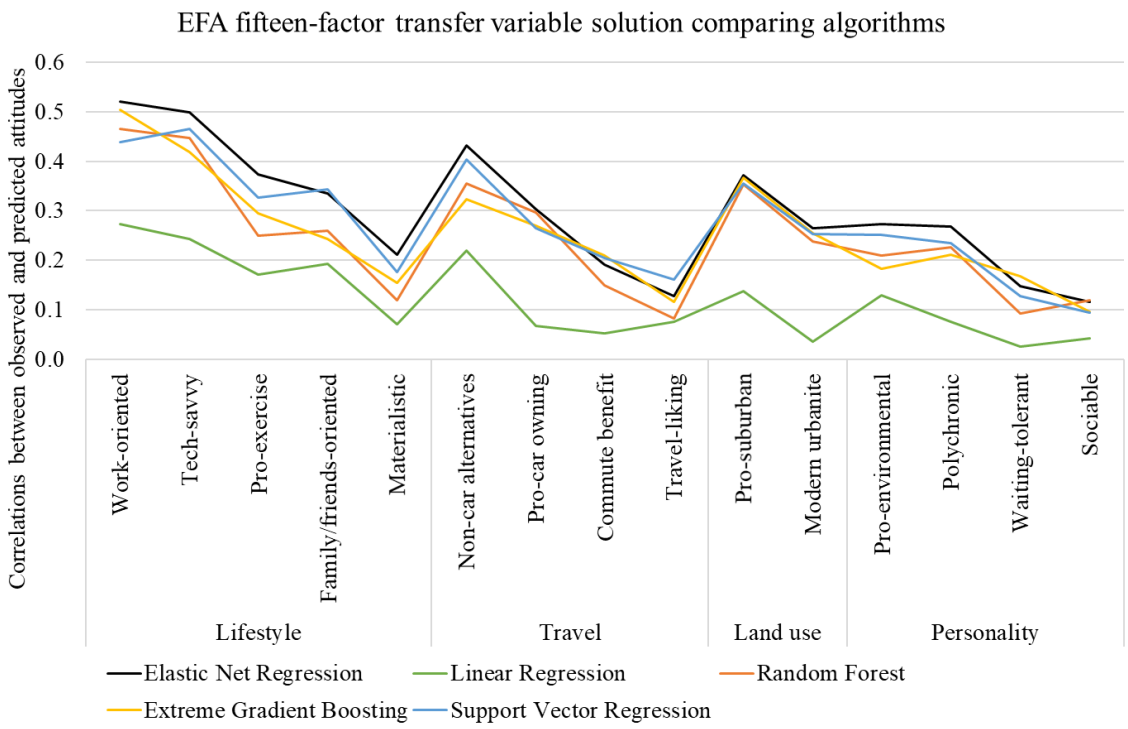

Figure 3.19. Comparison of performance across various algorithms

The comparisons shown and/or discussed throughout the results section (Section 3.5) represent just a small slice of what can be explored, and it is entirely possible that more sophisticated ML or deep learning algorithms and/or tuning approaches may yield better results than those shown in this chapter. Nonetheless, it is hoped that the work and steps shown provide a structure by which analysts can begin these explorations on their own datasets, while advancing and streamlining the techniques and approaches discussed here.

3.6 Discussion

This chapter provided the technical framework for a dataset enrichment approach, namely, the transfer-learning method (Section 3.4) alongside an advanced and detailed application of using this method to develop transfer algorithms for bringing attitudinal constructs from a variable-rich but smaller statewide research-oriented survey (GDOT

survey) into a larger but subject-constrained behavioral survey (Georgia subsample of NHTS). The level of detail reported throughout this chapter is intended to allow analysts to get a sense of the possibilities that could be explored using this framework. To close, an integrated look at the results of the application is provided, followed by a discussion of the limitations of this approach, and lastly a summary of key takeaways.

3.6.1 Results

Figure 3.20 provides a final comparison of the best ("cherrypicked"; see for instance Table 3.1 and Table 3.2) results for each of the three types of common variables examined, relative to the results when using the combination of these common variables determined to be the "best" set (i.e., a combined dataset of all native common variables, all TM variables, and dimension reduced subsets that explained 50% of the variance in "EPA SLD and All Transit" and ACS datasets). The TM variables alone perform best relative to the land use and SED variables alone, and in general using the overall subset yields only small improvements relative to using the TM subset alone. These results suggest that when possible, using a diverse, passive dataset such as targeted marketing data can improve the performance of the transfer learning framework. However, it also shows that in the absence of externally appended datasets like TM and land use, native SED common variables may also perform relatively well – at least for some of the transfer variables in this study. For the travel-liking, polychronic, sociable, and waiting-tolerant attitudes, SED variables alone performed even better than the best results from all of the variable sets combined – another potential consequence of the curse of dimensionality (see, e.g., Section 3.5.3.2.1). For the tech-savvy, work-oriented, materialistic, and pro-exercise attitudes, SED variables alone

performed essentially as well as the best results from all databases combined. These findings are of course specific to this particular application.

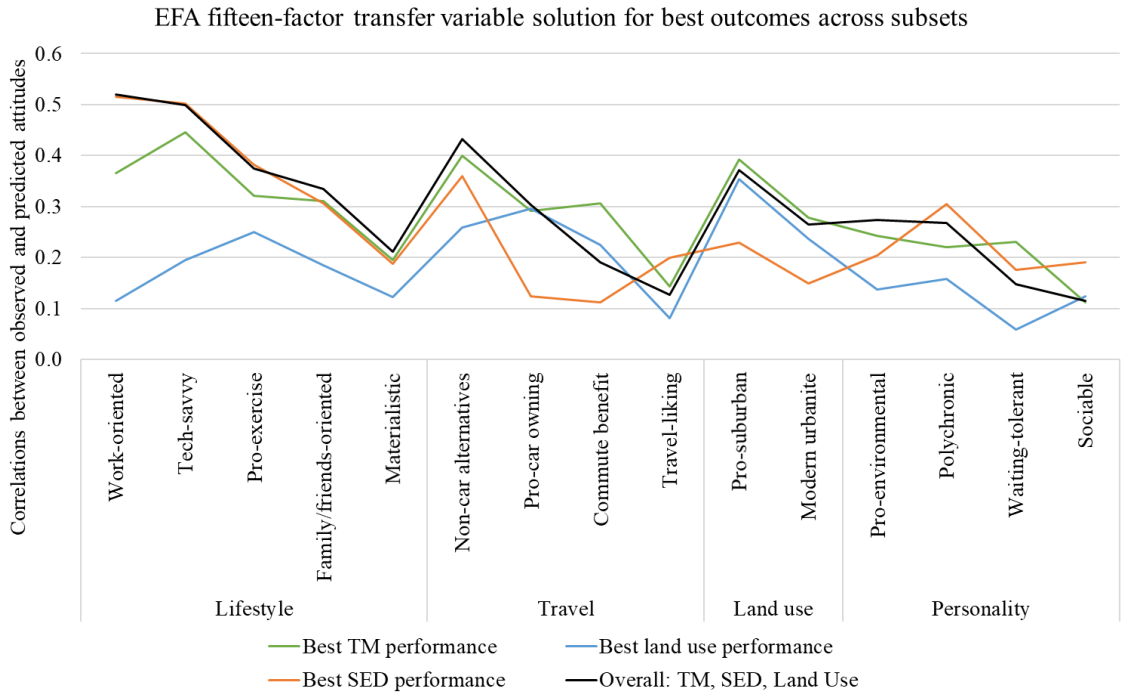

Figure 3.20. Comparison of best outcomes across subsets

In general, across the various combinations of inputs and algorithms examined over the course of the transfer process, the correlations between observed and predicted attitudinal constructs ranged from 0.1 to 0.5, with none of the domains (e.g., lifestyle, travel, etc.) outperforming the others decisively. While these internal validation numbers may at first appear to be less than ideal, they reflect expected outcomes for attitudinal variables which typically have R^2 (the square of the correlation) values of 0.1 to 0.2 (or lower) in the literature (see for example: Shaw, Malokin, Mokhtarian, & Circella, 2019). As a result, as noted before, the value of this transfer process is really best evaluated through external validation (Chapter 5): i.e., aiming to observe whether the imputed attitudes, as imperfect as they are, provide some benefit to the desired models.

3.6.2 Limitations

As with all survey data enrichment and data fusion methods, there are many caveats and limitations that should be kept in mind. The most critical of these are discussed here:

1. Firstly, ML transfer algorithms such as the ones applied in this chapter perform best as the size of the dataset increases. Typical transportation survey datasets that have transfer variables of interest tend to be relatively small, which in turn can decrease the stability and reproducibility of the transfer results obtained when using ML algorithms. For example, smaller datasets will be more prone to fluctuations depending on random variations in how the dataset is divided, and even more critically may not have enough cases to facilitate a training/test set split. Nonetheless, analysts may still apply the transfer learning framework on smaller datasets using non-ML algorithms.
2. Next, fulfilling the assumptions of spatial and temporal congruence as well as similarities in marginal distributions of key variables is difficult to achieve between datasets, regardless of field of study. It is important for researchers to examine and disseminate findings on the impacts of these incongruencies on transfer learning results in general.
3. Variables transferred through this process will have unique measurement errors as a result of the transfer process. This results in two sources of error that may have different scales in the enriched datasets. Additional work on quantifying this resultant measurement error and investigating the potential impacts on downstream analyses is needed. For example, while tools such as the Murphy-Topel correction exist for adjusting standard errors in predicted values that will be used in

downstream analyses, to our knowledge this hasn't been applied to ML algorithms (and/or isn't clearly accessible using available statistical software and packages; Murphy & Topel, 2002). One potential remedy is simply to be more conservative about the significance levels accepted in downstream analyses; this approach has not been implemented in this document but is certainly intended to be investigated in future work.

4. In addition, there may be conflicts/overlaps between the common variables used for the prediction of transfer variables and the variables that analysts wish to use downstream of the transfer process (i.e., in modeling efforts involving the resulting fused datasets). This overlap can also lead to ambiguity in what the resultant variable actually represents, thereby resulting in the reification fallacy for downstream analyses: for example, is the transferred variable simply a transformation of the input features? Future research could systematically investigate the impacts on downstream analyses of including various types of variables into the common variable pool. In particular, one potential advantage of the marker variable approach described in Chapter 4 is that it has the potential to transfer attitudes very effectively without (heavily, if at all) relying on other common variables that could be needed to perform "double duty" as explanatory variables in downstream models.
5. Along those lines, given that this process has the potential to introduce large errors and biases into the resulting fused datasets, the effectiveness of the variables transferred will lie in whether they provide any value to downstream applications. Thus, analysts using this method are encouraged to thoroughly validate their results

using internal and external validation procedures. Ultimately, it would be beneficial for many teams of transport and urban planning researchers to apply this method to bring various types of transfer variables into transport survey datasets, and thereafter to perform and share validation exercises that can help all analysts better understand the limitations of this approach in our field.

3.6.3 Key takeaways

To recap, transfer learning entails the use of common variables present between datasets to serve as inputs or features in algorithms or functions that are trained to predict output variables desired to be transferred across datasets. This chapter presented an advanced form of this framework by augmenting the common variable sets with passive and active data sources, and by using high performing algorithms to predict and transfer the variables of interest. Throughout Section 3.5, it was demonstrated that there are a seemingly infinite number of parameter combinations that can be varied within the three components of the framework (i.e., the transfer variables, features or common variables, and algorithms). Teams/analysts with enough resources may choose to follow an exhaustive process of determining the best feature subsets and algorithmic parameters to maximize the transfer performance. However, in the absence of the knowledge, time, and money necessary to apply an advanced and exhaustive form of transfer learning, – analysts can explore this method even in cases where limited common variables and algorithmic capabilities are present. Regardless of how simple or advanced the framework being applied may be, the ultimate assessment of usefulness lies in internal and external validation results, where is becomes necessary for the analyst(s) to assess the resultant performance based on their knowledge of the transfer variables. In the case of the application being shown in this chapter, it is known that attitudinal constructs specifically (and psychometric variables, in general) are among the most difficult variables to model, and as such, the internal validation results lie in the range of expected outcomes.

CHAPTER 4. ABBREVIATING SURVEY INSTRUMENTS USING MARKER VARIABLES

Survey designers increasingly have to balance breadth versus depth when designing survey instruments/questionnaires. This leads to tradeoffs that can result in a shortage of rich and potentially insightful variables available for use in survey analyses. The survey enrichment methods provided in Chapters 2 and 3 of this document are designed to address this problem by bringing variables into survey datasets from external datasets, be it passive (e.g., transactional data) or active sources (e.g., other surveys). This chapter explores an approach for obtaining some of this rich information directly within the survey (from the respondent), through the use of "marker statements" embedded within the survey instruments themselves. Marker statements are defined in this context as condensed/reduced sets of statements that are representative of a larger array of observed questions, variables, values, and/or developed constructs. This method of directly obtaining values from the respondents has the potential to reduce propagated error that can occur with the use of transfer-based enrichment approaches such as those discussed earlier. In line with the structure of prior chapters, this chapter provides:

1. a brief examination of the literature on abbreviating survey questionnaires;
2. a methodological framework for developing marker statements; and
3. an application of the method to develop and internally validate a marker set of attitudinal statements that could be integrated within future surveys.

In Chapter 5, the results of the application are externally validated using travel behavior models. However, note that in this case only external validation on the initial dataset is

possible because a test of the developed marker statements on a new survey has not yet been executed.

The work detailed in this chapter is from the following paper, which is currently in preparation:

Shaw, F. A. & Mokhtarian, P. L. (paper in preparation, available upon request from authors). An investigation into the development of psychometric markers that preserve explanatory power in travel behavior models.

4.1 Abstract

What if key information from rich, nuanced survey questions could be obtained using a fraction of the number of questions originally designed to capture this information? This would mean that survey designers could expand the breadth of surveys without significantly increasing survey lengths and thereby reducing response rates. This chapter presents one approach for achieving these goals by extracting reduced sets of statements representative of a larger array of questions (i.e., marker statements) for inclusion on survey instruments. After a presentation of the method, it is applied to 46 attitudinal statements in a statewide research-oriented transportation survey to develop a set of marker statements that could be integrated into future regional and national household travel surveys (e.g., NHTS) to obtain attitudinal variables typically not captured on travel surveys, and which are therefore seldom available for transport modeling. An internal validation of the framework applied to the GDOT dataset illustrates that across the attitudinal constructs, the extracted marker variables capture between 55 and 94% of variance present in the original attitudinal constructs. While variations of this method have

been used in other fields, it is not widely applied within transport and urban planning. It is hoped that the presentation of the method along with the sample application will encourage others to apply and further develop this method, or to validate the developed set of attitudinal marker statements identified in this chapter. Over time, this approach has the potential to significantly widen the pool of variables available for planning purposes, an outcome that could positively impact the infrastructure planning process in many ways (e.g., by bringing more user-centered variables into the transport planning process).

Keywords: survey design; survey instruments; survey questionnaires; survey length; short measures; machine learning; household travel survey; transportation survey; travel demand modeling; attitudes; attitudinal constructs; psychometric variables

4.2 Introduction

As discussed, it is the tradeoff of survey length relative to variable richness/survey breadth that often precludes surveys from capturing nuanced variables such as psychometric traits, which typically require extensive observations and/or questions on survey instruments. While smaller-scale, regional surveys may be able to include a more extensive range of questions relative to national, large-scale surveys, they too suffer from low response rates, a challenge that can be exacerbated by increased survey lengths. Accordingly, the method detailed here provides a framework for the development of marker statements that can reduce the number of questions needed for capturing nuanced variables, while still providing (some of) the rich information that would otherwise be lost by eliminating the entire set of related statements.

This chapter begins with a brief review of literature on abbreviating survey instruments and/or developing question banks that are intended to have the aim of the "marker statements" in this chapter (Section 4.3). The proposed methodological framework for extracting marker statements is presented in Section 4.4, followed by a sample application using the attitudinal questions in the GDOT survey to develop marker statements (Section 4.5). The chapter closes with a discussion of key takeaways and limitations (Section 4.6). The development of transport-relevant marker variables can facilitate the efficient acquisition of rich information within long-form travel behavior surveys, thereby facilitating improvement in model predictions without compromising survey response rates.

4.3 A review on abbreviated survey instruments

Given that a range of names and approaches have presumably been used across disciplines to achieve the same goal as the method discussed here, for clarity it is noted that combinations of the following search terms were used in conducting the literature review for this section: "abbreviated surveys", "short measures", and "scale abbreviation". After a general search on these terms, they were combined with transport-related keywords like "transport(ation)" and "urban planning". Due to the variety in terminology that likely exists outside of these terms (and of which we are not aware), it is acknowledged that this is not a comprehensive review. Such an undertaking could be useful for researchers across many disciplines. Based on the findings of this review, the development of approaches for shortening survey instruments appears to have been most commonly explored in the fields of marketing, psychology, and health, an unsurprising finding given the high prevalence of surveys and psychometric measures used in these fields (see for example: Aalto, Alho,

Halme, & Seppä, 2009; Kelly & Doriot, 2017; Marsh, Huppert, Donald, Horwood, & Sahdra, 2020; Rammstedt & John, 2007; Wassenaar et al., 2018). In many of these applications, the context and/or types of questions being abbreviated are very different from the applications within transport, and in fact utilize terminology and background outside the scope of transportation; and as such, are not extensively discussed here at this time.

4.3.1 Transport-related literature

On the other hand, a much smaller selection of literature within the transport field has explored the objective of shortening surveys/instruments. Cain et al. (2017) developed a 54-item abbreviated survey (MAPS-Abbreviated) from the original 120-item Microscale Audit of Pedestrian Streetscapes (MAPS); the items were selected from the original survey based on their correlations with physical activity, which was the dependent variable being modeled in the study. The team found that the MAPS-Abbreviated and original MAPS total scores had correlations of 0.94, with the abbreviated survey being related to physical activity outcomes similarly to the original MAPS questionnaire. Cerin, Saelens, Sallis, & Frank (2006) – from outside of transportation – developed the Abbreviated Neighborhood Walkability Scale (NEWS-A) using correlations between the NEWS-A and the Walk Score® index. Then, within transport, Silveira and Motl (2020) successfully validated the abbreviated version of the neighborhood environment walkability scale as an instrument that would provide perceived neighborhood walkability for individuals with multiple sclerosis. In addition to these efforts, there have been some proposals for a standard core set of attitudinal statements in transportation. One such effort applied discriminant analysis to identify the most powerful attitudinal questions that distinguish best between desired

segments of the population (Anable & Wright, 2013). In closing, the small sample of transport literature relevant to this method (of which the present author is aware), further motivates the framework and application detailed in this chapter.

4.3.2 Common methods/approaches

A range of methods have been used to extract reduced statements (or marker statements) for shorter instruments. Perhaps the simplest approach is the selection of items based on their correlations with a variable/outcome of interest or with the overall score on the questionnaire. Various types of correlations have been used for this purpose; examples include partial correlations and item-total correlations (Cain et al., 2017; Cerin et al., 2006; Kupper & Denollet, 2012). The second most widespread approach, particularly in the psychology domain, is the use of genetic algorithms (often considered a form of machine learning – i.e., automated pattern recognition) to find items that explain the most variability in the full measure (see for example: Basarkod, Sahdra, & Ciarrochi, 2018; Eisenbarth, Lilienfeld, & Yarkoni, 2015; Noetel, Ciarrochi, Sahdra, & Lonsdale, 2019; Sahdra, Ciarrochi, Parker, & Scrucca, 2016; Sandy, Gosling, & Koelkebeck, 2014; Yarkoni, 2010). The framework presented in this chapter is in line with the latter method, with the goal being to extract marker statements that explain the greatest amount of variability in the full set of statements (see Section 4.4.2.1 for more details) .

4.4 A review on abbreviated survey instruments

4.4.1 Overview of methodology

Figure 4.1 summarizes the methodological overview of the marker statement development process. To begin, there must exist a dataset (i.e., donor survey – D_D) that has already obtained a full set of observed statements that capture the information that is desirable for analysts to obtain on future surveys (i.e., recipient surveys – D_R), but for which a shortened version of these statements (i.e., marker statements) is necessary or desired due to space limitations or other constraints on the recipient survey(s). Thus, for simplicity as well as due to the parallel nature of the methodological process, the terminology between Chapter 3 and the one at hand (Chapter 4) is similar here whenever possible. The donor dataset input variables include common variables between donor and recipient datasets that are distinct from (but possibly correlated with) the marker variables of interest (X_D' and X_R'), as well as the common variable set that represents the marker statements themselves (Y_D' and Y_R').

Determining the marker common variables (Y_D' and Y_R') represents the first critical goal of this method – the marker statements are identical questions that are present in both the donor and recipient datasets and which will aid in bringing desired information into the recipient survey without necessitating the collection of the entire set of variables corresponding to Y_D (the entire set of variables corresponding to Y_D comprise the marker variables Y_D', as well as variables that contribute to the transfer variables of interest (Y_D''), but which are not part of the condensed marker set). Additional variables, X_D'' and X_R'', represent variables unique to the donor and recipient datasets respectively (i.e., not present

in the other dataset), but which are not of significance in this process. Given these definitions, the transfer process (τ) comprises a learning function $f(\cdot)$ that learns to predict Y_D based on X'_D and Y'_D. This learning function is then applied to X'_R and Y'_R to predict $\hat{Y}_R$. Thus, $Y_D = f_D(X'_D, Y'_D) + \varepsilon_D$, and $\hat{Y}_R = f_D(X'_R, Y'_R)$, where the learning function f_D is invariant between the donor and recipient domains. In comparing this approach to that shown in Section 3.4, it is seen that there are intentionally significant similarities, whereas the main difference is that there are now observed marker statements that are included in the set of common variables, thereby facilitating improved predictions of the desired information across datasets.

Note that it is not necessary to utilize additional common variables beyond the marker statements when expanding out the desired set of information. Further, it is possible that expanding out the marker statements into the full set of desired information Y_R may not be necessary, as the marker statements themselves (Y'_R) may provide enough information to fulfill the intended application. This would mean using the recipient dataset in the form shown in Figure 4.1, and not expanding out additional information from the marker statements. If the information *is* being expanded, then the assumptions and requirements associated with this method again are very similar to those discussed in Chapter 3, as we are once again using an algorithm developed on one dataset to aid in expanding data collected in another dataset. To recap briefly, the assumptions associated with the transfer process center on ensuring spatial and temporal congruence across the donor and recipient datasets, while also ensuring that the marginal distributions of common variables (most importantly for the SED variables) across datasets are not significantly different.

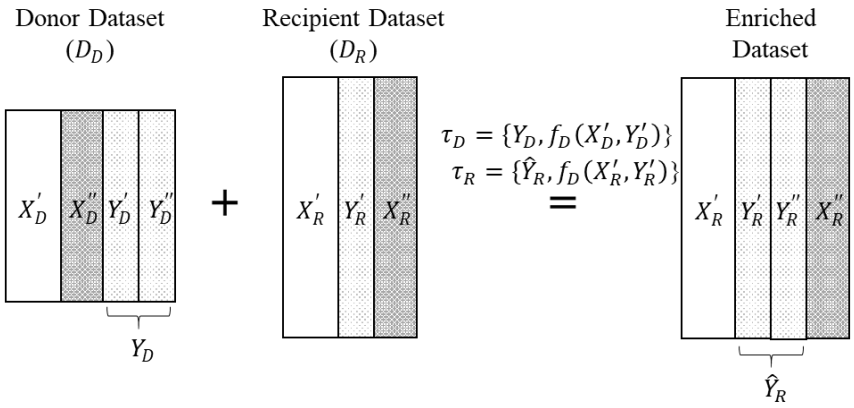

Figure 4.1 Overview of marker statements methodological process

4.4.2 Components of the process

In the following subsections, the primary three steps of the marker statement development process are detailed, showing aspects of each step that can be varied (see Figure 4.2). Note that this structure is at a higher/different level than the methodological overview presented in Chapter 3 because this process includes more components. For instance, the second step shown (i.e., marker statement utilization/expansion) encompasses all three components of the transfer process shown in Figure 3.3. Thus, the process developed in Chapter 3 essentially becomes a tool that is used within the framework discussed here. However, as introduced in the prior section, if the marker statements are *not* used to expand out the original statements or constructs, but rather just used directly, then step two in Figure 4.2 is omitted. In the case of the framework shown here, step two is discussed and applied.

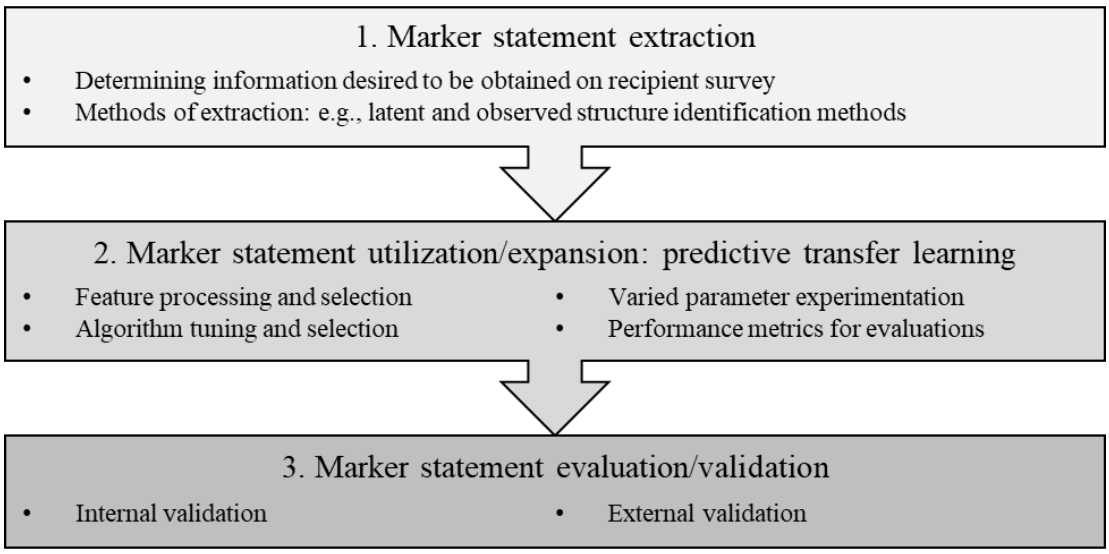

Figure 4.2 Components of marker statement development process

4.4.2.1 Extracting marker statements

The first step in the marker statement development process detailed here is to apply a statistical method that captures patterns of intercorrelation, covariance/variability/variance, and/or underlying structure amongst the full set of observed variables from which marker statements are to be extracted (see Figure 4.3). Virtually any method that captures interrelationships and patterns among a full candidate set of observed variables (Y_D) can be examined for this purpose; for instance, almost all latent (e.g., exploratory factor analysis) and observed structure identification methods (i.e., cluster analysis; dimension reduction methods like principal components analysis; automated pattern recognition methods like genetic algorithms, etc.) would be excellent contenders to explore.

Once the structural identification method of choice has been applied, the statements that are shown to capture/account for the largest amount of shared variance present in the corresponding clusters of statements (i.e., the bolded questions in Figure 4.3) represent the

pool of potential marker statements. The number of marker statements selected from the pool of potential statements may be constrained by the number of questions/space available on the survey instrument on which the marker statements will appear, as well as the desired set of statements, components(s), cluster(s), or construct(s) that the survey designer wishes to capture. Some software packages allow researchers to shorten questionnaires using item-cost parameters that capture the weight placed on having fewer items relative to having an instrument that explains more variance (Noetel et al., 2019).

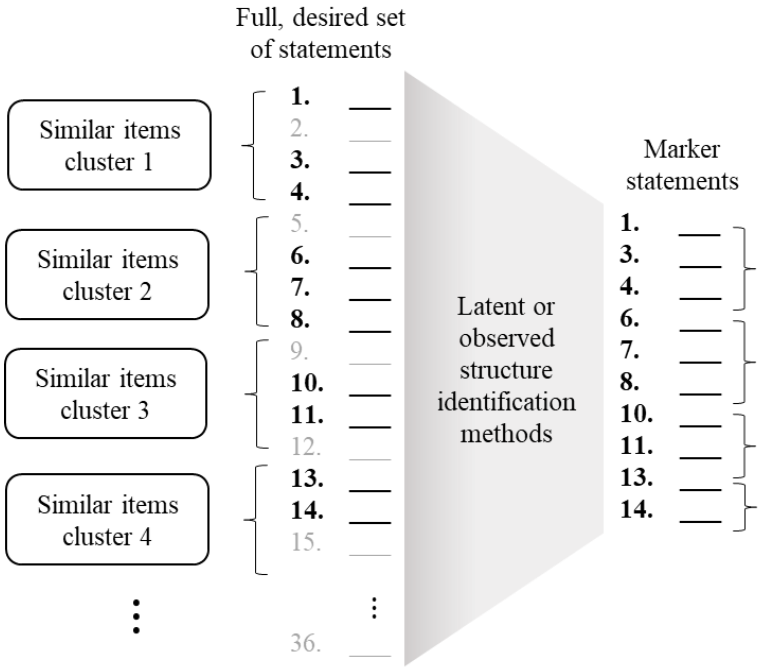

Figure 4.3. Extracting marker statements

4.4.2.2 Utilizing marker statement for survey instrument expansion

Once the marker statements have been selected from the full set of observed variables in the donor survey, the next step is to use the marker statements either: (a) directly, or (b) to transfer/impute the full set of information desired into the recipient survey, whether that full set of information be the original set of statements (Equations 3a

and 4a) or the underlying cluster(s) or construct(s) (Equations 4a and 4b; also see Figure 4.4). In this section, the latter use of marker statements is discussed. This stage of the marker statement development process encompasses the experimentation with and selection of the three components (transfer variable, algorithm, features) of the predictive transfer learning process discussed in Chapter 3, which as noted, becomes a tool of sorts within this process. To summarize the process verbally, as the equations show, the goal is to develop an algorithm predicting the observed full set of statements or constructs from the marker statements, for the donor dataset. This algorithm is then applied to the recipient dataset to obtain the information desired from the marker statements.

Firstly, analysts need to determine the form of the information desired to be brought into the recipient survey – i.e., is it preferred to bring in the full set of statements, or should constructs, components, or factors (i.e., reduced forms of the full statements) be transferred instead? With regards to the features/inputs, native and augmented common variables can be used in conjunction with the small set of common marker variables to improve the algorithm training and prediction process, thereby facilitating improved predictions of the full set of information desired. However, we emphasize that it is not necessary or required to use these additional (to the marker statements) CVs. Similarly, and again in line with Chapter 3, the algorithms selected for use may range from traditional regression or choice models to more advanced ML and deep learning algorithms. Analysts may examine a range of algorithms and their respective parameters before deciding on the algorithm that yields the best performance in this context.

$$Statements_{Donor} = f_{Donor}(Markers_{Donor}, CV_{Donor}, augCV_{Donor}) + \varepsilon_{Donor} \quad (3a)$$

$$Constructs_{Donor} = f_{Donor}(Markers_{Donor}, CV_{Donor}, augCV_{Donor}) + \varepsilon_{Donor} \quad (3b)$$

$$Statements_{Donor} = f_{Donor}(Markers_{Donor}, CV_{Donor}, augCV_{Donor}) + \varepsilon_{Donor} \quad (4a)$$

$$Constructs_{Donor} = f_{Donor}(Markers_{Donor}, CV_{Donor}, augCV_{Donor}) + \varepsilon_{Donor} \quad (4b)$$

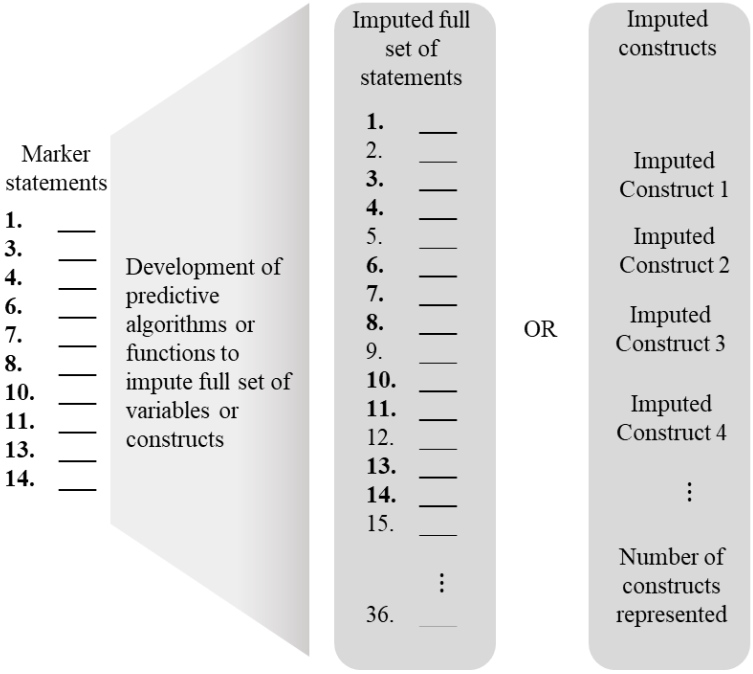

Figure 4.4. Utilizing extracted marker statements

4.4.2.3 Validating marker statements

As with all survey enrichment methods, there are two forms of validation that should be pursued – internal validation and external validation. In the literature, particularly from a psychometric standpoint, it is considered good form for analysts using abbreviated questionnaires to internally validate the shortened surveys using metrics like

reliability (internal consistency), test-retest reliability, content validity, factorial validity, criterion-related validity, etc. (Noetel et al., 2019; Sandy et al., 2014). Not all of these metrics will be applicable to all applications, as is the case in the one presented here; however, the terms are mentioned here because it is considered important for analysts using shortened questionnaires to investigate which of these measures may be applicable to their context(s). Following internal validation, analysts should externally validate the abbreviated information and/or the resultant predicted full measures by utilizing and comparing the information from the shorter survey relative to the full survey in domain-area specific models and applications.

4.5 Marker statement application

Having now provided some insight into how the marker statement development process could be partitioned and approached, this section details a sample application of the process. Specifically, the steps discussed in the prior section are applied to the rich array of 36 attitudinal statements present in the GDOT survey (i.e., the donor survey). Figure 4.5 provides an application-specific overview of the process, from marker statement identification to validation. For this application, the internal validation procedure mirrors that of Chapter 3, with the metric of choice being the examination of content validity using correlations of the marker statements and predicted constructs with the full set of observed information. The marker statements (Chapter 5) are externally validated using transport choice models, which allows for the quantification of differences in model fit and predictive power that occurs when using the full suite of attitudinal questions relative to the markers.

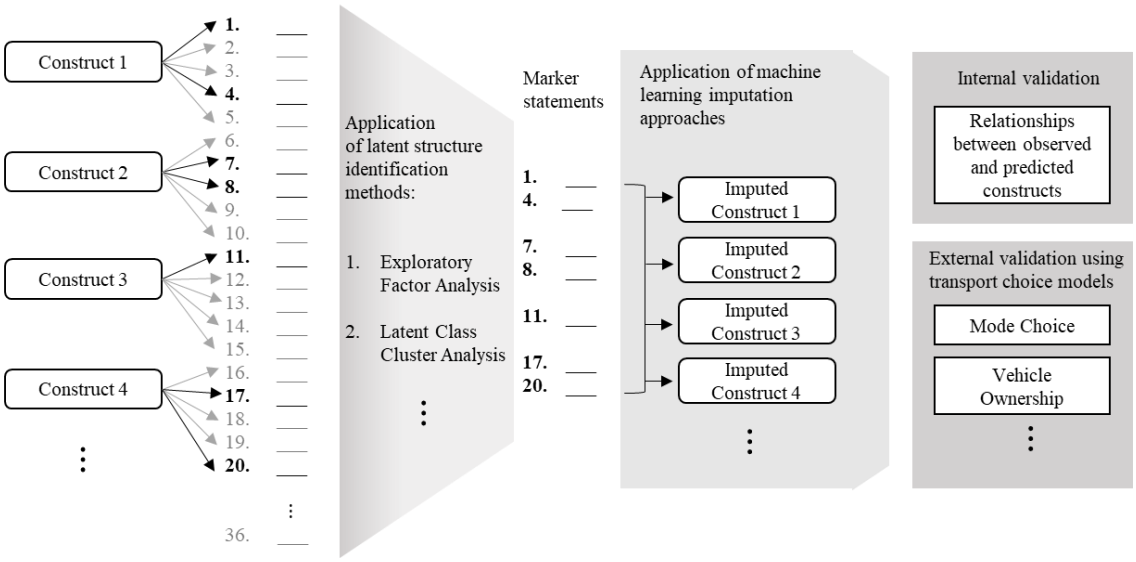

Figure 4.5. Application of marker statement development methodology
(as applied to latent attitudinal constructs)

4.5.1 Extracting marker statements

The first step in the process is to apply a structure identification method to the full set of observed variables. This can be done in various ways, but in this application, the latent structure identification method of exploratory factor analysis with oblique rotation (Oblimin, delta = 0) is used. Figure 4.6 provides a graphical overview of the variance partitioning approach used by exploratory factor analysis (EFA); in this method, the constructs are underlying latent variables that explain the shared variance among the observed variables. The full set of attitudinal statements with their resultant constructs are shown in Table 4.1. For consistency throughout the document, the EFA solution shown here is the same as that obtained in Chapter 3 when determining the latent attitudinal constructs for the transfer process. To recap, the eigenvalue-greater-than-one rule was applied to the initial eigenvalues, and thereafter, interpretability, simple structure,

communalities, and factor correlations were used to develop and tune the final solutions (Stevens, 2009).

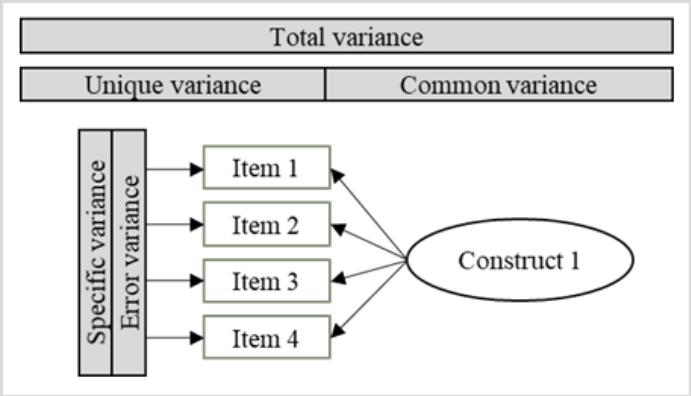

Figure 4.6. Visualization of exploratory factor analysis variance partitioning theory

The next step is to identify statements that have the highest loadings on each latent construct/factor. Note that only 36 statements are represented in this solution, as the others are not related to the 15 extracted constructs. The 15 developed constructs are deemed the full set of information that is desired on the recipient survey, and so there is no problem with the elimination of the 11 statements that are not influenced highly by these constructs. In Table 4.1, the statements that are being best explained by the latent constructs, and which therefore serve as the marker variables, are in bold text. For simplicity, we selected the statement with the highest pattern loading for each construct, thereby yielding 15 marker statements. However, there are an assortment of other approaches that could be used to obtain marker statements – for example, one might choose to retain all statements with loadings greater than 0.5 on the selected constructs. Each analyst may have different constraints on the number of marker statements that can be included on the recipient survey, and this will drive differences in extraction approach at this point in the process.

Table 4.1. Determining marker statements for attitudinal variables

Construct/ Factor	Statement	EFA Pattern Loadings[a]
Non-car alternatives	**s. I like the idea of walking as a means of travel for me.**	**0.730**
	ae. I like the idea of bicycling as a means of travel for me.	0.727
	c. I like the idea of public transit as a means of travel for me.	0.350
Tech-savvy	**g. Learning how to use new technologies is often frustrating for me.**	**-0.938**
	af. I am confident in my ability to use modern technologies.	0.835
Commute benefit	**y. My commute is a useful transition between home and work (or school).**	**0.693**
	q. My travel to/from work (or school) is usually pleasant.	0.610
	as. I wish I could instantly be at work (or school) – the trip itself is a waste of time.	-0.421
Modern urbanite	**l. I like the idea of having stores, restaurants, and offices mixed among the homes in my neighborhood.**	**0.432**
	k. My phone is so important to me, it's almost part of my body.	0.398
Work-oriented	d. At this stage of my life, having fun is more important to me than working hard.	-0.475
	u. I'm too busy to have as much leisure time as I'd like.	**0.675**
Materialistic	**ah. I usually go for the basic ("no-frills") option rather than paying more money for extras.**	**-0.598**
	n. The functionality of a car is more important to me than the status of its brand.	-0.451
	z. I would/do enjoy having a lot of luxury things.	0.417
	aq. I like to wait a while rather than being first to buy new products.	-0.364
	b. I prefer to minimize the amount of things I own.	-0.344
Polychronic	**ag. I prefer to do one thing at a time.**	**-0.919**
	e. I like to juggle two or more activities at the same time.	0.725
Pro-environmental	**v. Cost or convenience takes priority over environmental impacts (e.g. pollution) when I make my daily choices.**	**-0.941**
	ar. I am committed to an environmentally-friendly lifestyle.	0.550
Family/friends-oriented*	**p. Family/friends play a big role in how I schedule my time.**	**-0.602**
	w. It's okay to give up a lot of time with family and friends to achieve other worthy goals.	0.467
Pro-suburban	**aa. I prefer to live in a spacious home, even if it's farther from public transportation or many places I go to.**	**0.651**
	f. I see myself living long-term in a suburban or rural setting.	0.362
	z. I would/do enjoy having a lot of luxury things.	0.439
Waiting-tolerant*	**al. Having to wait is an annoying waste of time.**	**0.958**
	h. Having to wait can be a useful pause in a busy day.	-0.526
Travel-liking*	**ac. I generally enjoy the act of traveling itself.**	**-0.716**
	a. I like exploring new places.	-0.563
Sociable*	**x. I consider myself to be a sociable person.**	**-0.687**
	o. I'm uncomfortable being around people I don't know.	0.462
Pro-car-owning	**t. I definitely want to own a car.**	**0.882**
	j. I am fine with not owning a car, as long as I can use/rent one any time I need it.	-0.599
	ak. I like the idea of driving as a means of travel for me.	0.460
Pro-exercise	**ao. The importance of exercise is overrated.**	**0.756**
	m. I am committed to exercising regularly.	-0.702

[a]Note that in obliquely rotated factor analysis, pattern loadings represent regression coefficients for the factor model.
*The loadings on these statements must be reversed during interpretation.They have been reversed as needed in all model results shown

4.5.2 Utilizing and internally validating marker statements

In all results shown in this section, we used elastic net regression with hyperparameters tuned via grid search and an 80/20 training/test split. However, due to the small number of features when using just the marker statements, we find almost no difference in performance when using this complex formulation of elastic net regression relative to when using ordinary least squares (OLS) regression. Accordingly, this illustrates that more complex tools like machine learning algorithms may not be necessary when the common variable set is small. In addition, we note that because Chapter 3 extensively experiments with various parameters during the transfer process, the application shown throughout this section refrains from doing so. However, it is advised that should the analysts have the time and resources, the same approach taken in Chapter 3 should be applied to this process.

Figure 4.7 summarizes the correlations between observed and predicted attitudes (i.e., the internal validation metric of choice), when using marker statements as the common variable predictors relative to the other feature subsets explored in Chapter 3. The black line in the figure shows the results when using all 15 attitudinal marker statements to help impute each of the 15 attitudinal constructs. The blue line shows the results when using only the corresponding attitudinal marker statement for the respective construct. Both approaches to using the marker statements can be tested for any application.

It is not surprising that the imputation does least well when the marker variables have lower loadings on the associated original construct (e.g., modern urbanite and materialistic, with respective marker loadings of 0.432 and -0.598). It is more interesting

that many of the correlations between original and imputed constructs are rather high even when the associated marker statement has a lower loading on the original construct (e.g., the family/friends orientation marker statement has an original construct loading of -0.602, but the imputed construct has a correlation of more than 0.851 with the original construct). It is also seen that using only marker statements as common variables decidedly outperforms the use of any of the other subset(s) of common variables explored in Chapter 3. This hints at the potential power and simplicity of the marker variable approach relative to the extensive and complex process involved in assembling, pre-processing, and deploying the targeted marketing and land use feature sets. On the whole, it can be seen that marker statements are able to produce imputed construct scores that are very strongly correlated with the original scores for most of the constructs in this application. Be aware that this may not be the case for all applications – particularly those that may have lower correlations between the marker statements and the overall information being brought/transferred into the recipient survey.

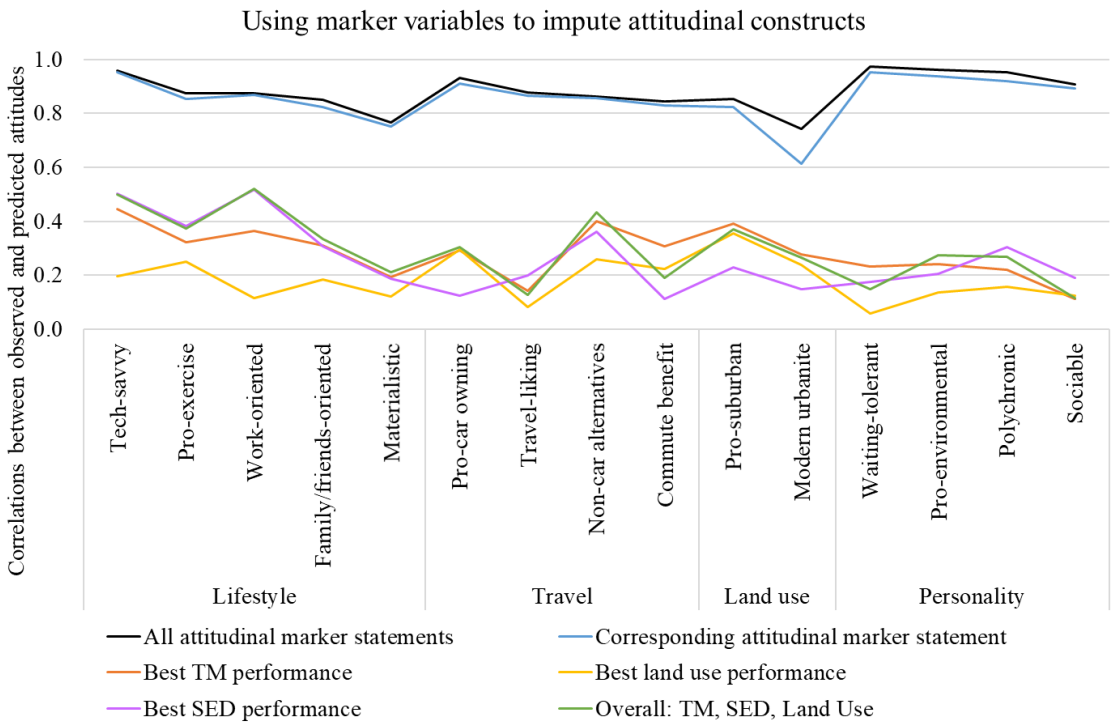

Figure 4.7. Comparison of performance across subsets, relative to using marker variables

In Figure 4.8, we test subsets of the 15 marker variables in accordance with the domain in which they are classified. The green line signifies the performance of the high performing combination subset of the native (SED) and augmented (land use and travel) common variables tested in Chapter 3, and is included on this chart for comparison purposes. We see that even when significantly decreasing the number of marker variables available and using out-of-domain marker variables, we can still obtain improved performance for 10 of the 15 attitudinal constructs (i.e., pro-exercise, materialistic, modern urbanite, travel-liking, commute benefit, pro-car owning, pro-environmental, waiting tolerant, polychronic, sociable) relative to when using a combined subset of SED variables, targeted marketing data, and land use data all together. This serves to confirm the power of marker variables, even when the number of marker statements is severely limited.

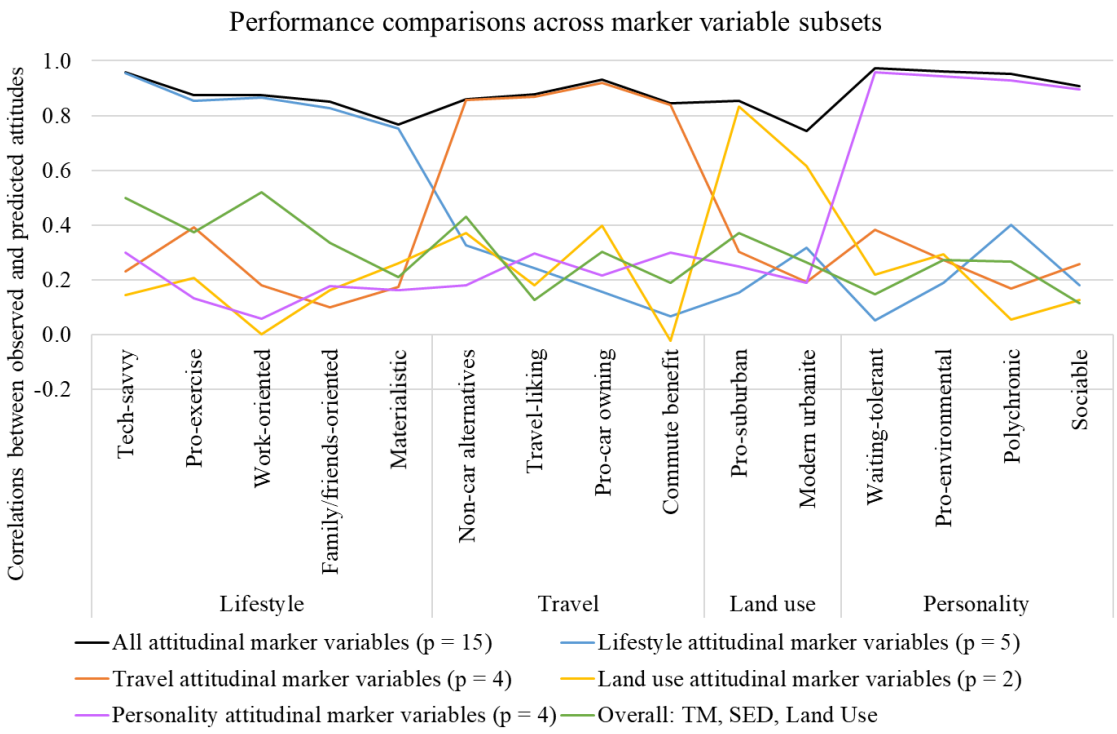

Figure 4.8. Comparison of performance across marker variable subsets

4.6 Discussion

This chapter detailed a framework for selecting a condensed set of attitudinal statements – marker statements – that facilitate expanding the number of diverse variables on surveys, without substantially increasing survey length and thus burden on respondents. The framework is built on the predictive transfer learning approach discussed in Chapter 3, and aims to bring information into recipient surveys based on desired information present in donor surveys. The primary difference is that in this approach, the analyst has obtained a small selection of highly relevant data points (related to the transfer information of interest) on the recipient survey that can help to improve the transfer of the desired variables. This approach has not been explicitly applied to a recipient dataset at this point and is intended for future work by the author.

As in Chapter 3 with the transfer learning framework, this approach assumes that the relationship (i.e., captured by transfer algorithm) between the marker statements and the imputed information is stable outside of the specific characteristics of the donor survey study. This may certainly be more-or-less true for certain domains of variables; however, in the case of the application shown here, the evolution of transport-related attitudes over time has not been studied extensively in the literature, and thus cannot be commented on from that perspective. However, in general there are known to be different types of attitudes – some are long term and inherent, while others are subject to environmental conditions and societal expectations (Maio & Haddock, 2009). Thus, it is important for the analyst to critically examine whether the assumptions in the marker variable approach affect the particular application of use. For additional discussion on the limitations that can be posed by incongruences across datasets, the reader is directed to Section 3.6.2 (and Chapter 3) for more details.

Despite the limitations, there are significant benefits that the marker variable approach can offer. Firstly, it is clear that marker variables have the potential to greatly improve the performance of the transfer process, thereby decreasing the amount of error that is introduced into the transferred information. Relatedly, the higher transfer performance for a small number of common marker variables (again, this is application-specific) may mean that advanced approaches like machine learning and the use of novel, big data are not necessary. This makes the approach more accessible to all transportation professionals. The marker variable approach can also reduce potential conflicts between the common variables used for the prediction of transfer variables and the variables that

analysts wish to use downstream from the transfer process (e.g., using transferred variables to explain/predict a variable that was already used in the transfer function).

In closing, in the field at hand, the development of a set of transport-relevant (in this case attitudinal) marker variables can facilitate the efficient acquisition of a rich set of information within long-form travel behavior surveys (for survey designers who see fit to include these marker statements in their survey instruments), while ensuring that survey response rates and model predictive power are not compromised.

CHAPTER 5. EXTERNAL VALIDATION OF ENRICHMENT VARIABLES

Having presented and applied various approaches for bringing new variables into survey datasets, the resulting enriched datasets must now be externally validated. Following directly from Chapters 2, 3, and 4, the enrichment or transferred variables that are externally validated in this section include: targeted marketing data (Section 5.1), attitudinal constructs transferred using SED, TM, and land use variables (Section 5.2), and attitudinal constructs transferred using attitudinal marker variables (Section 5.3). The external validation procedure for the enriched datasets entails examining the usefulness and value of the enrichment variables (relative to SED explanatory variables[2]) in modeling travel behavior outcomes such as vehicle ownership, ridesharing frequency, public transit usage, and so on (see Table 5.1). The metrics used for comparison across the travel behavior models are model fit, interpretability, and predictive accuracy. Whenever possible, the enrichment variables are modeled for both the GDOT (donor) survey and the NHTS (recipient survey), with the donor survey models serving as a benchmark of the value of the enrichment variables.

Table 5.1 summarizes the unweighted distributions of the travel behavior variables for the GDOT survey and NHTS datasets. As with the common ("across-survey") variables used in prior chapters, the travel behavior variables for external validation had to be harmonized between surveys (see Table D1 in Appendix D for more information). It can

[2] A list of the SED variables used in the models presented in this section can be found in Section 3.5.3.1.

be seen that the distributions/values for travel behaviors related to motor vehicle use: i.e., average vehicle miles driven, household vehicle ownership, and carshare usage, are similar across surveys. On the other hand, there are greater divergences in the distributions for ridesharing, public transit, and bicycle usage. Such divergences may be due to differences in question wording and/or answer categories. While the best effort was made to equate questions, it was not always (and will rarely be) possible to *perfectly* harmonize complex/non-traditional variables (i.e., non-SED variables) across surveys. In fact, simply finding similar, domain-specific variables on differing survey instruments is often difficult in-and-of itself, and this has often precluded external validation from being executed at all. Thus, slight differences in question harmonization should be kept in mind when analyzing the results, but do not undermine and should not prevent the external validation process from taking place.

Table 5.1 details the travel behavior variables for both the statewide *and* Atlanta region subsets of the GDOT and NHTS datasets (see Kim et al., 2019 for more information on the Atlanta-area subset). Both the statewide and Atlanta region subsets are examined in this section because travel services like public transit and ridesharing are more uniformly available in Atlanta (which results in a reduction in the skew of the response distributions for these behaviors), and this may conceptually and empirically improve the ability to model these behaviors using the available explanatory variables. One drawback of limiting some models to the Atlanta region is the accompanying reduction in the number of cases. Throughout the external validation models developed in this section, carsharing is not modeled for either survey due to the sparse occurrence of usage in both surveys, while public transit usage and bicycle frequency models are omitted for similar reasons for the

NHTS. Vehicle miles driven could not be harmonized across surveys (see Table C1 in Appendix C) and for this reason, is only reported here for the GDOT survey.

A range of external validation models are developed throughout this chapter. In the external validation for the targeted marketing data (Section 5.1) and attitudinal marker variables (Section 5.3), – only a naïve procedure that confirms the usefulness of the enrichment variables using simple regression models is provided, with more complex options discussed. On the other hand, for the attitudinal constructs transferred using SED, TM, and land use variables as features, an exhaustive approach is taken with the intention of demonstrating possible external validation paths that can be explored, and especially in the case when the transfer variables are psychometric and/or latent attributes (Section 5.2). A final note is that while it is recommended to split the data into training and test sets when developing external validation models, the sample size of the Atlanta region datasets is not optimal for this. Further, given that the focus here is on *relative* comparisons, not using training/test sets here will not change the final conclusions. Another factor influencing the decision to bypass this step is the range of model types and number of models developed throughout this section – adding additional formulations/comparisons would simply be overwhelming for both reader and author. Nonetheless, when the situation and sample size allow, the earlier position regarding training/test set use in this document is recommended.

Thus, as might be seen over the course of this discussion, there are many roadblocks and complexities that can threaten to derail external validation efforts. However, regardless of any drawbacks or challenges, analysts should prioritize executing external validation in whatever form possible or available, while acknowledging the (inevitable) caveats and

limitations clearly. External validation is critical to moving the field forward with regards to data enrichment and forecasting model improvement.

Table 5.1 Overview of travel behavior variables used for external validation

Variable	Categories	Georgia				Atlanta region			
		GDOT Survey[a] N = 2694		NHTS Survey[a] N = 4577		GDOT Survey[a] N = 878		NHTS Survey[a] N = 1349	
		N	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Household vehicle ownership[b]	0	64	2.38	185	4.04	16	1.82	41	3.04
	1	740	27.47	1368	29.89	235	26.77	428	31.73
	2	1047	38.86	1793	39.17	361	41.12	548	40.62
	3	524	19.45	778	17.00	168	19.13	217	16.09
	4	198	7.35	305	6.66	62	7.06	78	5.78
	5	71	2.64	97	2.12	18	2.05	24	1.78
	6+	50	1.86	51	1.11	18	2.05	13	0.96
Ridesharing usage frequency	Never used/no longer use	1847	68.56	4235	92.53	447	50.91	1143	84.73
	Less than once per month	501	18.60	106	2.31	223	25.40	59	4.37
	1-3 times a month	257	9.54	123	2.69	147	16.74	72	5.34
	1-2 times a week	54	2.00	83	1.81	41	4.67	57	4.23
	3 or more times a week	18	0.69	28	0.61	13	1.48	18	1.33
Vehicle miles driven[c]	Continuous variable	144.09 (142.45)		–	–	152.62 (141.84)		–	–
Public transit usage frequency	Never used/no longer use	2027	75.24	4198	91.72	551	62.76	1144	84.80
	Less than once per month	358	13.29	79	1.73	198	22.55	48	3.56
	1-3 times a month	101	3.75	120	2.62	71	8.09	64	4.74
	1-2 times a week	31	1.15	74	1.62	14	1.59	42	3.11
	3-4 times a week	26	0.97	61	1.33	16	1.82	29	2.15
	5 or more times a week	27	1.00	42	0.92	15	1.71	21	1.56
Bicycle usage frequency	Never used/no longer use	2050	76.10	4338	94.78	680	77.45	1282	95.03
	Less than once per month	311	11.54	0.00	0.00	99	11.28	0.00	0.00
	1-3 times a month	146	5.42	0.00	0.00	43	4.90	0.00	0.00
	1-2 times a week	69	2.56	154	3.36	26	2.96	44	3.26
	3-4 times a week	39	1.45	66	1.44	9	1.03	22	1.63
	5 or more times a week	43	1.60	18	0.39	10	1.14	1	0.07
Carshare usage frequency	Never used/no longer use	2560	95.03	4550	99.41	821	93.51	1339	99.26
	Less than once per month	76	2.82	14	0.31	34	3.87	4	0.30
	1-3 times a month	19	0.71	6	0.13	8	0.91	3	0.22
	1-2 times a week	8	0.30	2	0.04	2	0.23	0	0.00
	3 or more times a week	8	0.30	3	0.07	1	0.11	2	0.15

[a] Frequencies do not add up to 100% or the total N because of missing or not applicable cases/entries.
[b] The vehicle ownership models developed in this chapter use the continuous form of this variable, but the distribution by categories is shown in this table as it provides more information about the variable in the surveys.
[c] Mean (standard deviation) shown for vehicle miles driven. Equivalent measure not shown for NHTS because question could not be harmonized across surveys. See Table C1 in Appendix C for more information.

5.1 Targeted marketing data

As noted, the TM data that is externally validated in this section was integrated with the GDOT survey and NHTS datasets in Chapter 2. To assess the added explanatory power that TM variables can bring to travel behavior models, simple ordinary least squares (OLS) linear regression models for select travel behaviors from Table 5.1 are developed. For these models, SED characteristics and TM variables are entered simultaneously with no model refinement or pruning taking place as the intent here is simply to examine model performance rather than to focus on model interpretation. Further research on the external validation of TM variables may utilize more advanced approaches such as discrete choice models that allow for the collapse of the behavioral categories (e.g., binary logit and ordinal logit models), and machine learning models that allow for the use of feature importance metrics to identify the specific TM variables that contribute most to improving predictive accuracy. However, as before noted, due to more detailed external validation procedures shown for other enrichment variables later in this chapter, only a naïve external validation exploration for the TM data is shown here.

Figure 5.1 illustrates the improvement in model performance obtained when the explanatory variables include both SED characteristics and TM variables, relative to models that include only SED characteristics as explanatory variables. The dimension reduction method of PCA was used to develop sets of TM components that account for 50% (p = 97) of the variance present in the full, processed set of TM variables (p = 1128; see Section 3.5.3.2 for insight into how these variables are processed). The SED variables used are consistent with the set used in earlier chapters and contain a mix of individual and household-level variables (Section 3.5.3.1). The use of the TM principal components is

tested relative to SED variables only, as the latter are the primary explanatory variables available in national and regional household travel surveys. Across all of the models shown in Figure 5.1, the TM subset tested results in a nontrivial improvement to the adjusted R^2 values, relative to when using just SED variables as explanatory variables. The lift in adjusted R^2 values due to the TM components at times more than doubles the adjusted R^2 values seen when using just SED characteristics and is especially pronounced for the models of bicycle and public transit usage frequency, where SED variables have relatively poor explanatory power.

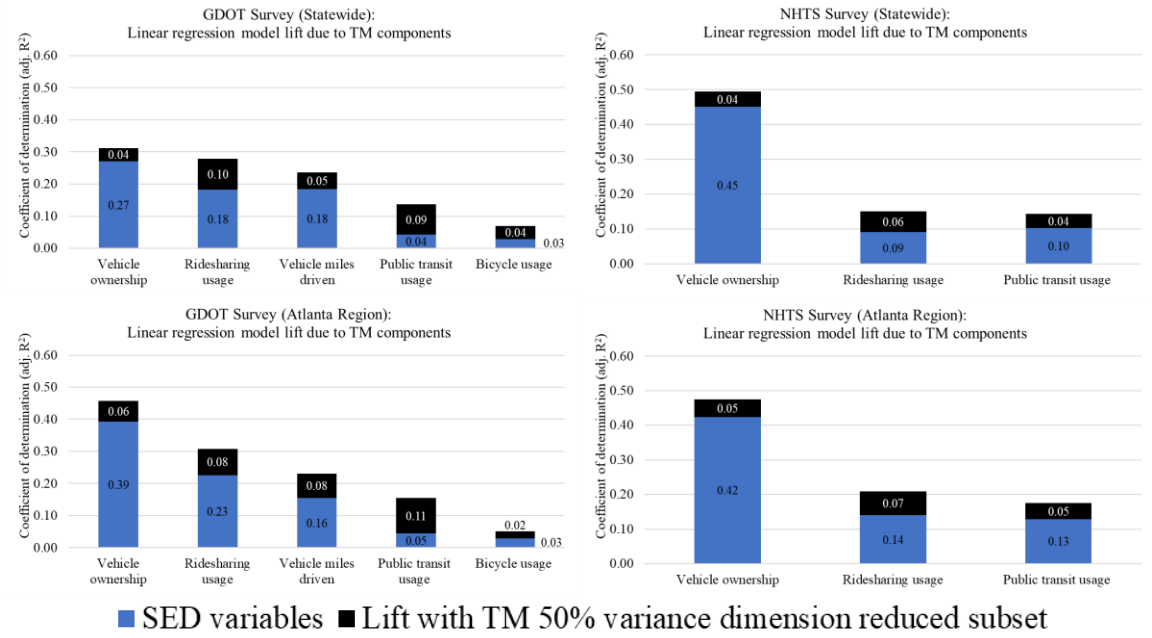

Figure 5.1. Linear regression travel behavior model lifts due to TM components

Figure 5.2 shows the performance of the TM components when they are entered independently of the SED characteristics (i.e., when they are the sole explanatory variables). It is seen that for ridesharing usage in both surveys, the use of the TM subset alone outperforms the SED subset. In the GDOT survey, this is also true for the public transit usage model. Overall, the results shown in Figure 5.1 and Figure 5.2 suggest the

potential value of TM variables in being able to improve model performance across a range of travel behavior models. Further investigations into the role of TM variables in travel demand modeling and forecasting are therefore warranted.

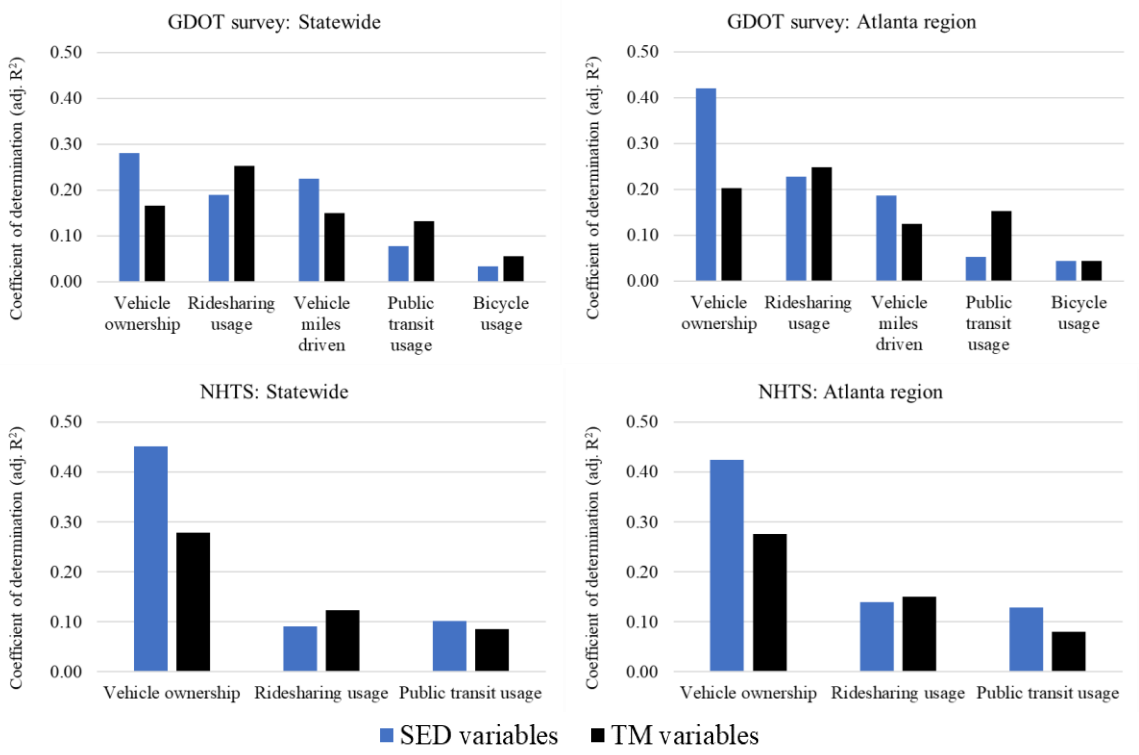

Figure 5.2. Linear regression travel behavior models with SED and TM variables independently included as explanatory variables for the GDOT survey and NHTS

5.2 Attitudinal transfer variables

The next task is to validate the attitudinal constructs that were transferred using a combination of SED, TM, and land use variables in Chapter 3. For these enrichment variables, linear regression, binary logit, and latent class choice models are used to test the usefulness and/or value of attitudinal variables for modeling select travel behaviors (chosen from those in Table 5.1). It is pertinent to note that in an application such as the one shown here, it is especially important that analysts explore the correlations between the variables transferred (i.e., the dependent variables) and explanatory variables used in downstream

applications such as external validation models. As in this case, there may be unavoidable overlap present between the features used to transfer the variables and the downstream explanatory variables (e.g., SED variables), and this could result in spurious coefficients or other model estimation problems. In external validation procedures where model interpretation is desired, it is recommended that standard procedures be followed for addressing highly correlated explanatory variables; for example, variance inflation factors (VIFs) and pairwise correlations can be examined and used to eliminate explanatory variables that may be competing with others in the model(s). For information on the correlations between the SED characteristics and predicted and observed attitudinal constructs in this section, see Figures D1, D2, and D3 in Appendix D. Further illustrating the potential issue discussed here, is the fact that the predicted attitudinal constructs have higher correlations with SED characteristics relative to the observed attitudinal constructs (which can be seen at a quick glance comparing Figures D1 and D2, with Figure D2 having a much darker color distribution which indicates higher correlations). It is for this reason that if interpretation is to be used as an external validation metric, it is recommended to carefully examine the correlations and to prune the model accordingly.

5.2.1 Regression models for travel behavior usage frequencies

First, to lay some initial groundwork, naïve linear regression models are developed to examine the effects of the attitudinal constructs developed based on the *observed* attitudinal indicators, as well as the *predicted/transferred* attitudinal constructs. The linear regression model formulation is consistent with the standard OLS regression model, and the explanatory variables entered comprise SED characteristics (Section 3.5.3.1) and attitudinal constructs (15 EFA constructs obtained from the best subsets of TM, SED, and

Land Use, using elastic net regression, as discussed in Section 3.6.1 and shown in Figure 3.20). For this section (i.e., 5.2.1), all SED variables and attitudes are used in the models with no model refinement/pruning taking place as the goal of the external validation at this level is to establish a basic understanding of the usefulness of attitudes, rather than to develop models for which coefficients will be interpreted and conclusions be drawn. The external validation models developed in 5.2.2 and 5.2.3 have been tuned and the resultant final models interpreted accordingly.

5.2.1.1 Varied travel behavior models

External validation regression models for a range of travel behaviors are summarized in Figure 5.3. The charts serve to confirm the general findings reported in the literature that attitudes, such as those obtained in the GDOT survey, can improve travel behavior models (Domarchi et al., 2008; Kuppam et al., 1999; Mokhtarian & Salomon, 1997). As seen in Figure 5.3, both types of attitudinal constructs yield improvements to the model fits shown (adjusted R^2), relative to when using just SED variables as explanatory variables. For the NHTS data relative to the GDOT survey data, smaller improvements to total model lift are observed as a result of the transferred attitudes; however, due to differences in initial model fit, the improvement percentages across both the GDOT and NHTS models are similar (i.e., on the order of 30%). Of the three travel behaviors with the overall best model fits (i.e., after including all variables) – vehicle ownership, ridesharing frequency, and vehicle miles driven – the ridesharing model illustrated the largest improvement in model fit (both in magnitude and percentage/ratio) for both datasets with the introduction of attitudinal constructs. As a result, the regression models for ridesharing are discussed further in Section 5.2.1.2.

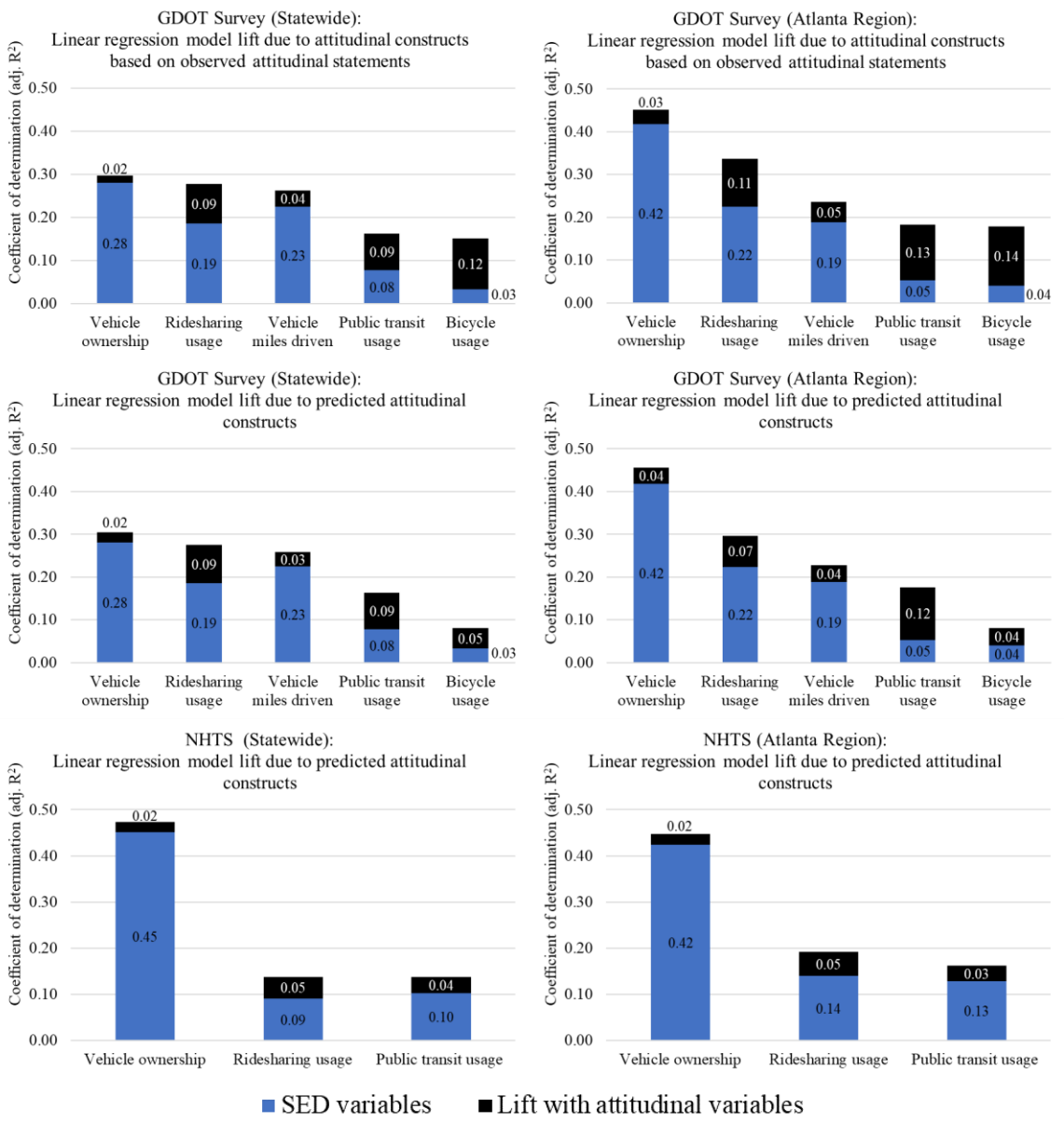

Figure 5.3. Linear regression travel behavior model lifts due to observed and predicted attitudes

It is important to note that for the GDOT dataset, as shown in Figure 5.4, when the attitudinal constructs are included as the sole explanatory variables in these travel behavior models, the transferred/predicted attitudinal constructs tend to outperform the constructs based on observed indicators (with the exception of the bicycle frequency outcome variable). This could potentially be due to the correlations between the predicted attitudinal

constructs and key excluded explanatory variables (like SED characteristics; see correlation tables in Appendix C) or may be attributable to unobserved nuances or value that the transfer process and/or algorithm is able to capture and bring to the predicted constructs. For vehicle ownership, the predicted attitudes in the NHTS outperformed all model formulations for GDOT; however, the opposite is true for ridesharing and public transit usage. This may be due to survey differences in question formulation and distributions as before mentioned.

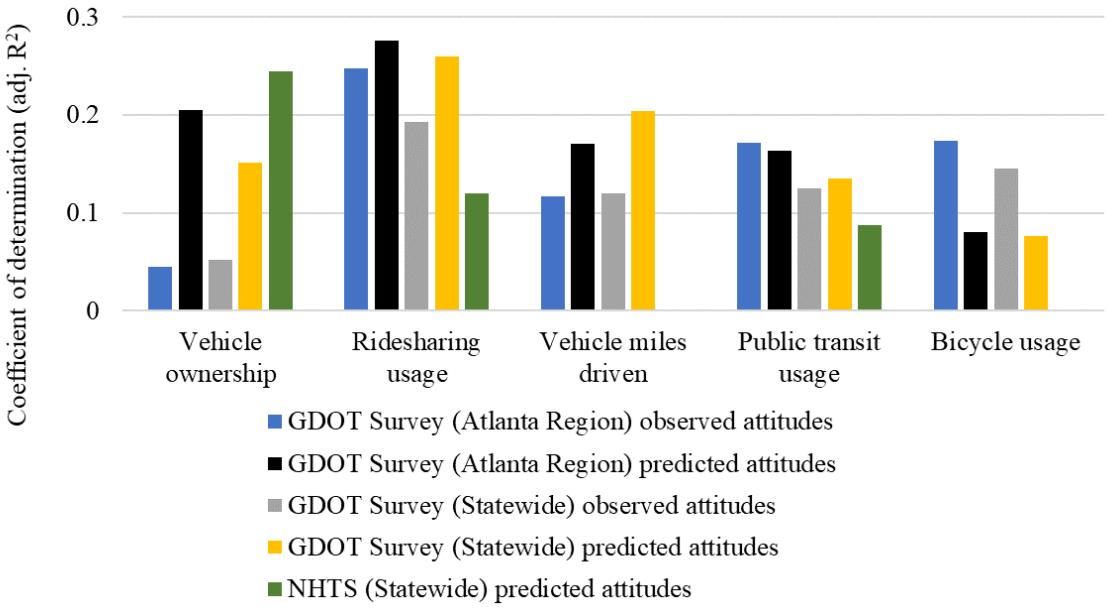

Figure 5.4. Linear regression travel behavior models with only observed or transferred attitudinal constructs as explanatory variables for the GDOT survey and NHTS

5.2.1.2 Regression models for ridesharing usage

After exploring the effects of attitudes on multiple travel behavior outcomes, ridesharing behavior is selected for further examination in the more detailed external validation efforts shown next. Ridesharing usage is selected for the reasons already stated in the prior section, as well as because for the GDOT survey, the ridesharing regression

model has the second greatest model lift (in magnitude) – after public transit – with the introduction of attitudes, but has more active users than public transit (see Table 5.1 and Figure 5.3). Meanwhile, for the NHTS, the ridesharing regression model has the largest model lift (in magnitude and ratio) for both the statewide and Atlanta-area models. Note that these observations and findings are distinctive to the region being studied; for example, the metropolitan Atlanta area has a very limited public transit system and areas outside of Atlanta are less likely to have widespread ridesharing services available. For these reasons, the follow-up external validation models focus on ridesharing within the *Atlanta region*. Figure 5.5 re-summarizes the model fits for the ridesharing linear regression models shown in this section, confirming that the introduction of the transferred attitudinal constructs yields a ~35% improvement in model fit for both surveys.

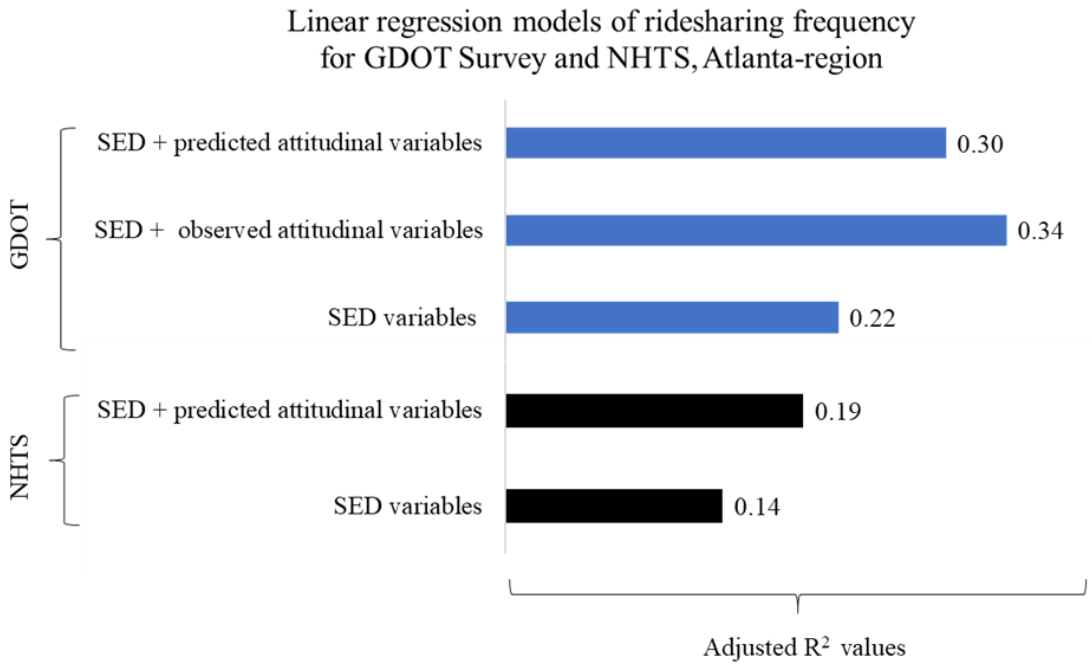

Figure 5.5. Linear regression modeling of ridesharing usage

5.2.2 Discrete choice models for ridesharing usage

As discussed in the preceding section, after examining the effects of attitudinal constructs on modeling a range of travel behaviors, ridesharing usage is selected as the dependent variable of choice to be examined in further external validation efforts which include traditional binary logit models as well as latent class choice models. Given the low occurrence of ridesharing usage frequencies in some response categories (see Table 5.1), the outcomes were collapsed into two options: ridesharing users and non-users, and binary choice models are utilized. In the event that the distribution of the behavior being modeled is more robust and multiple categories are retained, ordinal and/or multinomial choice models may be used in place of binary choice models. Prediction accuracies and model fit (McFadden's pseudo r-squared (ρ_{EL}^2) values) are the primary metrics used to compare models during the external validation process for ridesharing usage. Furthermore, these models are refined at a more thorough level that those in the prior section, and as such can also be examined and compared for interpretative value.

The work detailed in this section is from the following paper, which is currently in preparation:

Shaw, F. A., Etezady, A., & Mokhtarian, P. L. (paper in preparation, available upon request from authors). An investigation into the effects of observed and imputed attitudinal variables on ridesharing behavior: assessing the value of machine learning-based attitudes imputed into the Georgia NHTS subsample.

5.2.2.1 Binary logit models

The binary logit model (BLM) of ridesharing adoption developed for this section follows the standard formulation:

$$Y_i^* = \beta_i' X_i + \varepsilon_i ;$$

$$Y_i = 1 \; if \; Y_i^* \geq 0; \text{otherwise } Y_i = 0,$$

where Y_i^* is an unobserved, continuous latent variable that represents the propensity of each person i to use ridesharing, while its unstarred counterpart, Y_i, is the binary variable indicating whether ridesharing was used or not. X_i is the vector of characteristics (SED variables and attitudinal constructs) that are hypothesized to influence ridesharing usage for each individual, while β_i' is the vector of unknown but to-be-estimated coefficients that reflect the effects of the associated X_i characteristics on Y_i^*. Finally, the error term ε_i captures the influence of unobserved variables on the associated outcome Y_i^*. As with the regression models developed earlier, the explanatory variables (X_i) comprise SED and attitudinal constructs. However, here the SED variables are entered first and variables with high correlations and/or VIF values are removed, followed by variables that are insignificant. Next, the attitudinal variables are added to the retained SED variables, and the same model refinement process is repeated. In essence, this process assesses whether the attitudinal variables offer any additional explanatory power after a "best SED-only" set of explanatory variables has been identified.

5.2.2.1.1 Georgia Department of Transportation (GDOT) Survey

Table 5.2 summarizes the BLMs of ridesharing adoption developed for the Atlanta-region subset of the GDOT survey. The adjusted ρ_{EL}^2 metric of model fit indicates that for

the GDOT dataset, as seen with the linear regression models, the best model performance for ridesharing use occurs when the attitudinal constructs based on *observed* indicators are included in the model alongside SED variables. Specifically, improvements of ~33% in model fit are seen relative to the SED-only model when including predicted or transferred attitudes, whereas improvements of ~47% are seen when including attitudinal constructs based on observed indicators. Given the imperfect reproduction of observed attitudes by the transfer function (see, e.g., the correlations between observed and predicted attitudes shown in Figure 3.20 of Section 3.6.1), this is not surprising. This suggests that obtaining the attitudes themselves from respondents may be of greater value and supports the use of the marker statement methodology discussed in Chapter 4. Nonetheless, even though the observed constructs do better than the predicted ones, the 33% improvement in model fit offered by the latter (over the SED-only model) is still very desirable and suggests that the predicted attitudes have significant potential/value to offer.

Similarly, from an interpretation perspective, nine attitudinal constructs based on *observed* indicators are significant with logical coefficients, relative to the four *predicted/transferred* attitudinal constructs that are significant in the respective model. Again, this is consistent with the attenuation (perhaps into insignificance) of coefficient estimates that often characterizes variables measured with error (Stevens, 2009). Furthermore, some of the SED variables become insignificant in the model that includes predicted attitudes, which is likely attributable to the use of SED characteristics as input features during the transfer learning process used to predict the attitudinal constructs (see Chapter 3 for more information).

Table 5.2 Binary logit models of ridesharing adoption for GDOT survey, Atlanta[a]

	GDOT Binary Logit SED-only	GDOT Binary Logit SED & Predicted Atts	GDOT Binary Logit SED + Observed Atts
	N = 866	**N = 866**	**N = 866**
Predictors	*Coefficient*	*Coefficient*	*Coefficient*
Intercept	1.26**	0.04	1.08*
SED			
Education	0.23***	--	0.21**
Age	-0.05***	-0.03***	-0.05***
HH income	0.39***	0.30**	0.27***
HH size	-0.18**	--	-0.14*
HH vehicles	-0.13*	--	--
Attitudes			
Commute benefit	NA	--	-0.20**
Urbanite	NA	1.75***	0.37***
Materialistic	NA	--	0.35***
Pro-exercise	NA	--	0.18**
Pro-suburban	NA	-1.33***	-0.41***
Waiting-tolerant	NA	-1.74***	-0.22**
Travel-liking	NA	2.41***	0.22***
Sociable	NA	--	0.30***
Pro-car-owning	NA	--	-0.16*
Non-car alternatives	NA	--	--
Work-oriented	NA	--	--
Tech savvy	NA	--	--
Family/friends-oriented	NA	--	--
Pro-environmental	NA	--	--
Polychronic	NA	--	--
Fit measures			
$\mathcal{L}(\mathbf{0})$	-600.26	-285.56	-934.99
$\mathcal{L}(\boldsymbol{c})$	-599.93	-285.58	-576.53
$\mathcal{L}(\widehat{\boldsymbol{\beta}})$	-502.15	-230.94	-452.51
ρ^2_{EL} ($\mathcal{L}(\mathbf{0})$ base)	0.16	0.19	0.52
Adjusted ρ^2_{EL} ($\mathcal{L}(\mathbf{0})$ base)	**0.15**	0.17	0.51
ρ^2_{MS} ($\mathcal{L}(\boldsymbol{MS})$ base)	0.16	0.19	0.22
BIC	1044.89	491.99	955.46
AIC	1016.31	471.89	919.01

***, **, * = significant at 1% (0.01), 5% (0.05),10% (0.1), respectively.

[a]The market share of ridesharing users in the GDOT survey, Atlanta region dataset is 48.61% (i.e., 421/866).

Next, success tables and prediction accuracies are examined across the GDOT binary logit models to get a sense of the tangible improvement in predictions that attitudes provide when classifying ridesharing usage behavior (Table 5.3). First, some background information on success tables is provided here. The number of cases in the mn^{th} cell of a success table is calculated as follows:

$$N_{mn} = \sum_i I_i^m \hat{p}_i^n,$$

where N_{mn} is the number of cases whose observed choice is m and whose predicted choice is n. I_i^m equals 1 when the observed choice of case i corresponds to m, and equals 0 otherwise; and $\hat{p}_i^n$ represents the assigned probability (from the model being evaluated) for case i to choose n. In this case, m and n correspond to those who use ridesharing and those who do not (ridesharing users and non-users). In the unit-weighted success table, $\hat{p}_i^n$ is equal to 1 for the highest-probability choice (and 0 otherwise), while in the probability-weighted success table, $\hat{p}_i^n$ is the predicted probability that the choice models yield. Both types of success tables are provided so as to offer two points of comparison for other analysts. Many software programs automatically provide unit-weighted success tables and prediction accuracies, but cases have been made in the literature for the superiority of probability-weighted success tables in providing more accurate measures(Kim & Mokhtarian, 2018). The bolded values on the diagonals of the success tables represent cases that have been correctly predicted, while the off-diagonal values represent misclassified cases.

Table 5.3 summarizes the unit- and probability-weighted success tables and resulting prediction accuracies for the GDOT models of ridesharing usage developed and presented in Table 5.2. The prediction accuracies are calculated by summing the diagonal elements and dividing by the total number of cases (i.e., the sum of all four numbers in the success table). It can be seen that the introduction of attitudinal constructs yields improvements in the prediction accuracies of between 3 to 5 percentage points relative to models with SED characteristics only as explanatory variables. While these are modest improvements, they still serve to confirm the value that psychometric variables like attitudes may be able to improve modeling efforts in various ways. Perhaps more important than the improvement in predictive accuracy per se is the fact that the inclusion of attitudes makes the model more responsive to variables that are important to choice. In general applications of this approach to attitudinal as well as other "missing" variables, this could suggest new avenues for influencing behavior toward more sustainable choices, and improve planners' ability to predict the response to those new influences.

Table 5.3 Unit and probability-weighted success tables for GDOT binary logit models

	GDOT binary logit SED-only model			
	Unit-weighted		Probability-weighted	
	Predicted ridesharing users	Predicted ridesharing non-users	Predicted ridesharing users	Predicted ridesharing non-users
Observed ridesharing users	**285**	136	**249.92**	171.08
Observed ridesharing non-users	129	**316**	171.08	**273.92**
Prediction accuracy	**69.40%**		**60.50%**	
	GDOT binary logit SED and observed attitudes model			
	Unit-weighted		Probability-weighted	
	Predicted ridesharing users	Predicted ridesharing non-users	Predicted ridesharing users	Predicted ridesharing non-users
Observed ridesharing users	**293**	128	**269.73**	151.27
Observed ridesharing non-users	110	**335**	151.27	**293.73**
Prediction accuracy	**72.52%**		**65.06%**	
	GDOT binary logit SED and predicted attitudes model			
	Unit-weighted		Probability-weighted	
	Predicted ridesharing users	Predicted ridesharing non-users	Predicted ridesharing users	Predicted ridesharing non-users
Observed ridesharing users	**300**	121	**261.55**	159.45
Observed ridesharing non-users	110	**335**	159.45	**285.55**
Prediction accuracy	**73.33%**		**63.18%**	

5.2.2.1.2 National Household Travel Survey

Table 5.4 summarizes four BLMs of ridesharing adoption developed for the Atlanta-region subset of the NHTS. These models include the original NHTS data as well as a downsampled subset. Because the distribution of NHTS ridesharing users (15.27%) and non-users (84.73%) is significantly skewed relative to the GDOT data (48.38% and 51.62%, respectively), the downsampling of non-users to yield more balanced distributions is investigated as a potential fix. Downsampling is one of many approaches that can be

used to address severely imbalanced outcome distributions and produce models based on more comparable/manageable market shares (Google Developers, 2020). In the NHTS downsampled subset used in this section, a random sample of non-users is selected so as to reproduce the almost 50/50 market segmentation seen for the GDOT survey. The objective in reproducing the GDOT market share is so that the external validation models developed across surveys could be comparable. Another approach may be to randomly (down)sample non-users with the intent of simply correcting the skew slightly, for instance in this case by producing a 75/25 (non-users/users) market segmentation relative to the 82/18 segmentation present natively in the data. This latter approach was also tested for this chapter, but ultimately not shown due to the large number of model formulations already included.

The fit measures summarized in Table 5.4 indicate that, in the full NHTS sample, the predicted attitudinal constructs serve to improve the model fit by ~8% (adjusted ρ^2_{EL} of 0.51 versus 0.47) when included alongside SED characteristics. This is substantially less than the improvement seen in the (untuned) regression models developed on NHTS data earlier (~36%), as well as the improvement seen for the GDOT survey data (~33%). For the downsampled subset, however, the improvement is a similar 35% (adjusted ρ^2_{EL} of 0.23 versus 0.17). When using the market-share model rather than the equally-likely model as a benchmark, which is a better way to directly compare the two pairs of models to each other, we see similar improvements in both cases: 37% (0.16 versus 0.22) for the full sample, and 32% (0.25 versus 0.19) for the downsampled subset.

Further, compared to the refined binary logit model for the GDOT survey, there are slightly fewer predicted attitudinal constructs that are significant in the final model.

Nonetheless, it is believed that the predicted attitudes that are retained do serve to provide insight that could not have been obtained through the use of SED characteristics only. For example, those who have preferences for non-car modes of transport like walking, bicycling, and public transit are more likely to be ridesharing users, which is consistent with the literature that shows the importance of ridehailing and ridesharing as the last-mile connectors or needed on-demand transport options that are utilized by active transport and public transit mode users (Alemi, Circella, Mokhtarian, & Handy, 2019). It is also noted (in the full sample) that the importance of sociability to the adoption of ridehailing is an interesting insight that we have not seen elsewhere in the literature.

Table 5.4 Binary logit models of ridesharing adoption for NHTS, Atlanta region

	NHTS Binary Logit[a] **SED**	**NHTS Downsampled Binary Logit[b]** **SED**	**NHTS Binary Logit[a]** **SED + Predicted Attitudes**	**NHTS Downsampled Binary Logit[b]** **SED + Predicted Attitudes**
	N = 1349	**N = 412**	**N = 1349**	**N = 412**
Predictors	*Coefficient*	*Coefficient*	*Coefficient*	*Coefficient*
Intercept	0.98	2.72***	-0.46	1.05*
SED				
Education	--	--	--	--
Age	-0.06***	-0.06***	-0.05***	-0.04***
HH income	0.42***	0.42***	0.22***	0.25**
HH size	-0.28***	-0.25**	--	--
HH vehicles	-0.34***	-0.37**	-0.24**	-0.30**
Attitudes				
Commute benefit	NA	NA	--	--
Urbanite	NA	NA	--	--
Materialistic	NA	NA	--	--
Pro-exercise	NA	NA	--	--
Pro-suburban	NA	NA	-0.96**	-1.34**
Waiting-tolerant	NA	NA	--	--

Table 5.4 cont'd

Travel-liking	NA	NA	--	--
Sociable	NA	NA	1.84**	--
Pro-car-owning	NA	NA	--	--
Non-car alternatives	NA	NA	1.80***	1.96**
Work-oriented	NA	NA	--	--
Tech savvy	NA	NA	--	--
Family/friends-oriented	NA	NA	--	--
Pro-environmental	NA	NA	--	--
Polychronic	NA	NA	--	--
Fit measures				
$\mathcal{L}(\mathbf{0})$	-934.99	-285.56	-934.99	-285.56
$\mathcal{L}(\boldsymbol{c})$	-576.53	-285.58	-576.53	-285.58
$\mathcal{L}(\widehat{\boldsymbol{\beta}})$	-483.71	-230.94	-452.51	-215.05
ρ^2_{EL} ($\mathcal{L}(\mathbf{0})$ base)	0.48	0.19	0.52	0.25
Adjusted ρ^2_{EL} ($\mathcal{L}(\mathbf{0})$ base)	0.47	0.17	0.51	0.23
ρ^2_{MS} ($\mathcal{L}(\boldsymbol{MS})$ base)	0.16	0.19	0.22	0.25
BIC	1003.46	491.99	955.46	466.22
AIC	977.42	471.89	919.01	442.10

***, **, * = significant at 1% (0.01), 5% (0.05),10% (0.1), respectively.
[a]The market share of ridesharing users in the NHTS, Atlanta region dataset is 18.02% (i.e., 206/1143).
[b]The market share of ridesharing users in the NHTS, Atlanta region downsampled dataset is 50.00% (i.e., 206/412).

Table 5.5 summarizes the unit- and probability-weighted success tables and resulting prediction accuracies for the NHTS models presented in Table 5.4 (see Section 5.2.2.1.1 for background on success tables and prediction accuracies). It is seen that the introduction of the predicted attitudinal constructs yields improvements in prediction accuracies of between 2 to 3 percentage points relative to models with SED characteristics only as explanatory variables. Further, the downsampled models have lower prediction accuracies than the models developed on the original NHTS distributions. This is to be expected, as reductions in the skew of the distributions make it more difficult for models to predict the correct value simply based on the market share alone (i.e., "by-chance").

Note as well that the prediction accuracies for the downsampled data are similar in magnitude to those developed for the GDOT survey data. Thus, the metrics for the downsampled data may be said to provide a more realistic look at the performance of the models. On the other hand, assuming that the skewed distribution of the full-sample model is the more accurate one, it is noteworthy that the addition of attitudes disproportionally improves the prediction of the less-often chosen alternative, i.e., ridesharing: the SED-only model correctly predicts 28% of observed ridesharers to have chosen it (probability-weighted results), while the addition of attitudes allows the correct prediction of 32% of observed adopters (the comparable numbers for the non-adopters are 87% and 88%). Such improvements, though small, may be very useful for the analysis of seldom-chosen alternatives.

Table 5.5 Unit and probability-weighted success tables for NHTS binary logit models

	NHTS full-sample binary logit SED-only model			
	Unit-weighted		Probability-weighted	
	Predicted ridesharing users	Predicted ridesharing non-users	Predicted ridesharing users	Predicted ridesharing non-users
Observed ridesharing users	**30**	176	**57.44**	148.56
Observed ridesharing non-users	18	**1125**	148.56	**994.44**
Prediction accuracy	**85.61%**		**77.97%**	
	NHTS full-sample binary logit SED and predicted attitudes model			
	Unit-weighted		Probability-weighted	
	Predicted ridesharing users	Predicted ridesharing non-users	Predicted ridesharing users	Predicted ridesharing non-users
Observed ridesharing users	**44**	162	**65.94**	140.06
Observed ridesharing non-users	30	**1113**	140.06	**1002.94**
Prediction accuracy	**85.77%**		**79.23%**	
	NHTS downsampled binary logit SED model			
	Unit-weighted		Probability-weighted	
	Predicted ridesharing users	Predicted ridesharing non-users	Predicted ridesharing users	Predicted ridesharing non-users
Observed ridesharing users	151	55	128.04	77.96
Observed ridesharing non-users	60	146	77.96	128.04
Prediction accuracy	**72.09%**		**62.16%**	
	NHTS downsampled binary logit SED and predicted attitudes model			
	Unit-weighted		Probability-weighted	
	Predicted ridesharing users	Predicted ridesharing non-users	Predicted ridesharing users	Predicted ridesharing non-users
Observed ridesharing users	**159**	47	**134.50**	71.50
Observed ridesharing non-users	56	**150**	71.50	**134.50**
Prediction accuracy	**75%**		**65.29%**	

5.2.2.2 Latent class choice models

Having examined the usefulness and value of attitudinal constructs in modeling ridesharing using regression and binary choice models, the final and most advanced class of models tested for external validation purposes comprises latent class choice models (LCCMs). LCCMs are designed to capture unobserved heterogeneity in the population; and psychometric variables/traits such as the attitudinal constructs transferred in this study are considered to be well-suited for being able to delineate this unobserved heterogeneity through the formation of latent classes or subgroups. For reasons discussed already, ridesharing usage is the outcome being modeled here.

Figure 5.6 provides a visual overview of the LCCM model as applied to ridesharing usage, when the attitudinal constructs are based on observed attitudinal indicators (e.g., the GDOT survey attitudinal indicators). Figure 5.7 provides a visual overview of the LCCM model as applied to ridesharing usage when the attitudinal constructs are transferred using the process developed and shown in Chapter 3 (e.g., for the NHTS). LCCM simultaneously estimates the probabilities of each individual belonging to a specific latent class or group (membership model), and the probability of each individual making the choices being modeled conditioned on their membership in a specific class/group (choice model). For simplicity, in this application the attitudinal variables are limited to the class membership model, but in principle they could appear in the choice model instead or as well. For the SED-only LCCM formulations, no explanatory variables are entered into the membership model, which means the latent classes are based on the heterogeneity present in the outcome variable.

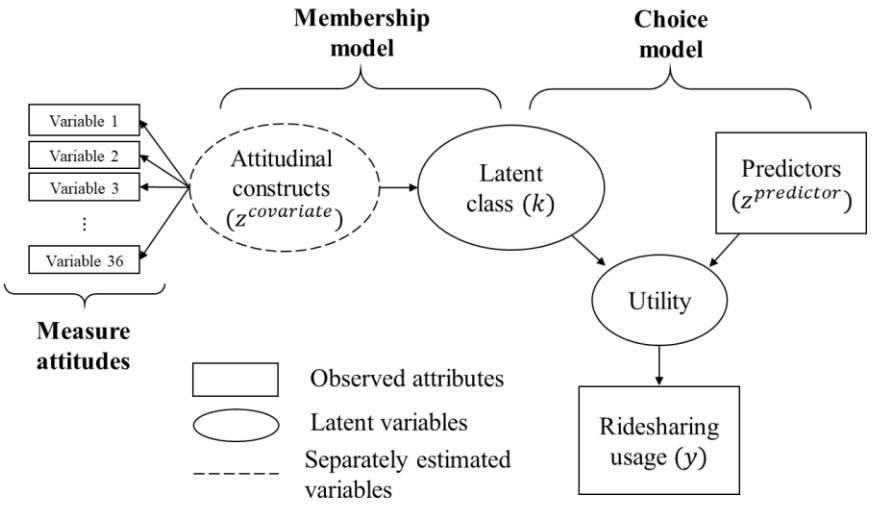

Figure 5.6. Latent class choice model of ridesharing usage when attitudinal constructs are based on observed attitudinal predictors

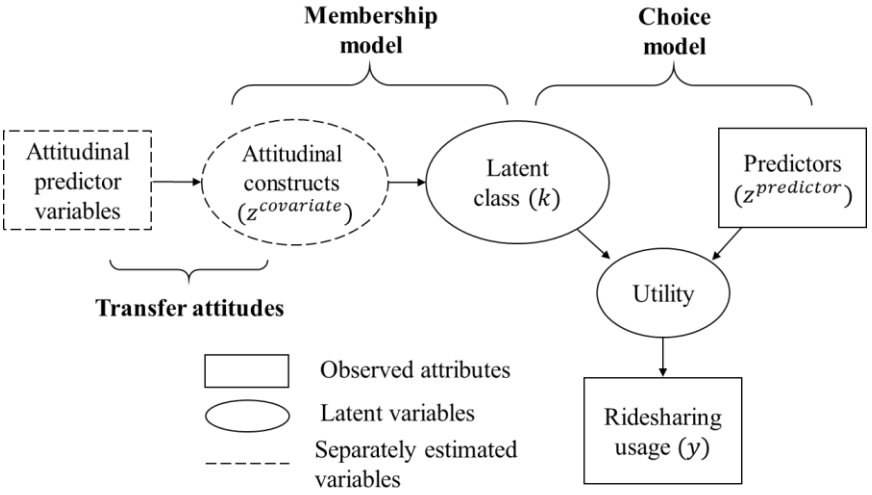

Figure 5.7. Latent class choice model of ridesharing usage when attitudinal constructs are transferred across surveys

The latent class choice model formulation utilized in this chapter is as follows,

$$P\left(\boldsymbol{y}_i \middle| \boldsymbol{z}_i^{covariate}, \boldsymbol{z}_i^{predictor}\right) = \sum_{k=1}^{K} P\left(k \middle| \boldsymbol{z}_i^{covariate}\right) P\left(\boldsymbol{y}_i \middle| k, \boldsymbol{z}_i^{predictor}\right),$$

where $\boldsymbol{y}_i$ represents the vector of outcomes being modeled (ridesharing usage in this case), $\boldsymbol{z}_i^{cov}$ represents the vector of covariates (attitudinal constructs), $\boldsymbol{z}_i^{pred}$ represents the vector of predictors (SED characteristics), and *K* represents the total number of latent classes. This can be expressed verbally as: the probability of observing a vector of outcomes given covariates and predictors, is equal to the membership probability for latent class k given the covariates, *multiplied* by the conditional probability of *y* given the class membership and predictors. As with the binary logit models, the LCCM models are refined and developed systematically with covariates and predictors entered one at a time, and multiple class solutions examined and compared before settling on the final models presented here. This approach ensures that model interpretation can be executed.

5.2.2.2.1 Georgia Department of Transportation (GDOT) Survey

In keeping with the external validation procedure thus far, three GDOT LCCM models are developed: a base model with SED characteristics only, a model with SED and predicted attitudinal constructs, and a model with SED characteristics and attitudinal constructs developed using observed indicators. By virtue of the definition of LCCMs, each LCCM model can entail the assumption of various numbers of classes, each of which has its own coefficients. As such, showing the full model for all classes across multiple models would be an overload of information that could distract from the purpose of this section – i.e., to evaluate the usefulness and value of predicted attitudes in modeling travel behavior, specifically ridesharing usage in this case. Thus, only the fit statistics across all three best models are summarized in Table 5.6, while the coefficients and detailed model information

for the best performing LCCM model that uses SED characteristics and predicted attitudinal constructs is shown in Table 5.7 and Table 5.8.

To reiterate, Table 5.6 summarizes the fit statistics across the best performing LCCM models using the three sets of explanatory variables explored throughout this validation process. The metrics present in Table 5.6 that should be compared across models are the adjusted ρ^2 values and the predictive accuracies. The predictive accuracies are developed using the process detailed in Section 5.2.2.2.1. As with all models, there are tradeoffs between fit and predictive accuracy that occur, and that were observed when developing the models shown here. However, given that the purpose of these models is for external validation, it is more important that similar approaches be taken across models (i.e., consistency), rather than to facilitate a discussion of which approach is better for this specific application. Of further consideration is the fact that in LCCMs there may be heterogeneities that could be observed by allowing more classes, thus yielding interpretive value but detracting from model fit or predictive accuracy.

After testing many models with varying numbers of classes, the two-class solution was selected as it was best able to balance trade-offs among model fit, predictive accuracy, and interpretation. Table 5.6 shows that, as with the binary logit models, the GDOT LCCMs that utilize attitudinal constructs perform better than a model without attitudinal constructs. However, conversely to the regression and BLM findings, the LCCM with *predicted* attitudinal constructs had almost equivalent model fits and predicted accuracies relative to the LCCM with attitudinal constructs developed based on *observed* indicator variables. Overall, the model fit (adjusted ρ_{EL}^2) and prediction accuracies observed for the LCCM models are similar in magnitude to those seen for the binary logit models (i.e.,

model fits of ~0.20 and prediction accuracies of ~68-73%), suggesting that the more advanced LCCM model specification is not bringing significant additional value with regards to model *performance*; however, it is in the discussion of interpretation that follows in which LCCM is expected to show its true power.

Table 5.6 Comparing LCCM models of ridesharing adoption for GDOT survey, Atlanta region

	GDOT LCCM SED-only	**GDOT LCCM SED + Pred. Atts**	**GDOT LCCM SED + Obs. Atts**
	N = 866	**N = 866**	**N = 866**
Fit measures			
Number of classes	2	2	2
Number of parameters	11	17	20
$\mathcal{L}(\mathbf{0})$	-600.26	-600.26	-600.26
$\mathcal{L}(\boldsymbol{c})$	-599.93	-599.93	-599.93
$\mathcal{L}(\widehat{\boldsymbol{\beta}})$	-503.44	-463.70	-458.71
ρ^2_{EL} ($\mathcal{L}(\mathbf{0})$ base)	0.16	0.23	0.24
Adjusted ρ^2_{EL} ($\mathcal{L}(\mathbf{0})$ base)	0.14	0.20	0.20
ρ^2_{MS} ($\mathcal{L}(\boldsymbol{MS})$ base)	0.16	0.23	0.24
BIC	1081.28	1042.39	1052.70
AIC	1028.88	961.40	957.42
Unit-weighted predictive accuracy (%)	67.32	73.67	72.06
Probability-weighted predictive accuracy (%)	59.57	63.90	63.96

Having discussed the fit statistics across the three GDOT models, model interpretation is now performed for the GDOT LCCM model with SED characteristics and predicted attitudinal constructs. Table 5.7 details the coefficients and Table 5.8 summarizes the segment specific shares/means for the model. Based on the profiles of the two emergent classes, the first class, which comprises 42.6% of the sample, is named "younger, tech savvy, ridesharing likely" given that 96% of the class report using ridesharing and 16% of the class fall into the youngest age category included in the survey (i.e., that of Millennials).

The latter point is significant because as shown in the survey descriptive statistics (see Table A1 in Appendix A), the GDOT survey is heavily skewed toward older individuals. Individuals in this class also have very positive attitudes toward technology. The second class, which comprises over 50% of the sample, is titled "older, pro-car, ridesharing hesitant" as only 13% of individuals in this class report using ridesharing. Further, only 13% of this class fall into the Millennials category. Individuals in this class have negative attitudes toward technology and are more positive toward owning cars and suburban living relative to individuals in the prior class. Thus, from an interpretation standpoint, the predicted attitudinal variables help to add additional insight into understanding the choices made with respect to ridesharing.

A final point to be noted here is that despite the overall prediction accuracies being similar in value to the BLMs developed in the preceding section, the LCCM yields a sizable segment (larger than the predicted ridesharing users correctly identified in the BLM) that is almost certain to be ridesharing users, and the SED and attitudinal profile of this segment is known (Table 5.8). In other words, the LCCM model shows an impressive ability to sort between those likely to be ridesharing users and non-users. This information has the potential to be extremely useful for market and policy analyses and predictions; and as before mentioned, could also provide avenues forward for influencing behavior toward more sustainable choices.

Table 5.7 Coefficients of GDOT LCCM with SED characteristics and predicted attitudes[a]

	Younger, tech savvy, ridesharing-likely (42.6%)		Older, pro-car, ridesharing hesitant (57.4%)	
	Coefficient	*p-value*	*Coefficient*	*p-value*
Outcome model (predictors)				
Intercept	-4.828	0.13	0.148	0.78
Education	2.160	0.07	-0.116	0.33
Millennials	-0.396	0.62	1.488	0.004
Low HH income (<$50K)	-2.170	0.11	-0.285	0.22
HH size	0.407	0.15	-0.338	0.035
Membership model (covariates)				
Intercept	-0.636	<0.001	0.636	<0.001
Tech savvy	1.069	<0.001	-1.069	<0.001
Materialistic	0.943	0.01	-0.943	0.01
Family/friends-oriented	-0.636	0.02	0.636	0.02
Pro-car-owning	-0.719	0.07	0.719	0.07
Pro-suburban	-0.574	0.04	0.574	0.04
Sociable	0.931	0.02	-0.931	0.02

[a]Note that effect coding has been used in the estimation of this model.

Table 5.8 Segment-specific shares/means of predictors and covariates for GDOT LCCM model

Variable	Younger tech savvy ridesharing-likely (42.6%)	Older pro-car ridesharing hesitant (57.4%)
	Variable *means/share* per class	Variable *means/share* per class
Outcome variable		
Ridesharing usage* (binary)	0.960	0.135
Predictors		
Education †	4.096	3.681
Millennials* (binary)	0.162	0.035
Low HH income (<$50K) * (binary)	0.139	0.313
HH size†	2.425	2.136
Covariates†		
Tech savvy	0.387	-0.069
Materialistic	0.173	0.008

Table 5.8 cont'd

Family/friends-oriented	-0.037	-0.007
Pro-car-owning	-0.182	-0.030
Pro-suburban	-0.209	-0.058
Sociable	0.096	0.039

† Segment-specific means
*Segment-specific shares (e.g., proportion of segment that can be classified as Millennials)

5.2.2.2.2 National Household Travel Survey (NHTS)

The tables included here (Table 5.9, Table 5.10, and Table 5.11) parallel those in the preceding GDOT survey LCCM section. As before, it can be seen that the introduction of predicted attitudinal constructs results in moderate improvements to the model fit and predictive accuracy. The NHTS downsampled model is selected for further interpretation as it is more comparable to the GDOT model discussed with regards to market shares and consequently, model performance.

Table 5.9 LCCM models of ridesharing adoption for NHTS, Atlanta region

	NHTS LCCM SED	NHTS Downsampled LCCM SED	NHTS LCCM SED + Predicted Attitudes	NHTS Downsampled LCCM SED + Predicted Attitudes
	N = 1349	**N = 412**	**N = 1349**	**N = 412**
Fit measures				
Number of classes	2	2	2	2
Number of parameters	13	11	14	14
$\mathcal{L}(\mathbf{0})$	-934.99	-285.56	-934.99	-285.56
$\mathcal{L}(\boldsymbol{c})$	-576.53	-285.58	-576.53	-285.58
$\mathcal{L}(\widehat{\boldsymbol{\beta}})$	-484.71	-220.18	-454.20	-202.25
ρ^2_{EL} ($\mathcal{L}(\mathbf{0})$ base)	0.48	0.23	0.51	0.29
Adjusted ρ^2_{EL} ($\mathcal{L}(\mathbf{0})$ base)	0.47	0.19	0.50	0.24
ρ^2_{MS} ($\mathcal{L}(\boldsymbol{MS})$ base)	0.16	0.23	0.21	0.29
BIC	1063.11	506.59	1009.29	488.79
AIC	995.42	462.36	936.39	432.49
Unit-weighted predictive accuracy (%)	85.10	74.76	85.32	76.70
Probability-weighted predictive accuracy (%)	77.86	64.14	79.03	67.55

Model interpretation is now examined for the downsampled NHTS LCCM model with SED characteristics and predicted attitudinal constructs. Table 5.10 details the coefficients and Table 5.11 summarizes the segment specific shares/means. The names of the classes are able to be kept almost consistent between the NHTS and GDOT LCCM models. The first class, which comprises 53.9% of the sample, is the "tech savvy, ridesharing-likely" class. Eighty percent of the class reports using ridesharing and as before, individuals in this class also have very positive attitudes toward technology. The second class, which comprises over 46.1% of the sample, is the "older, pro-car, ridesharing hesitant" class. For this model, 15 percent of individuals in this class report using ridesharing, and again, individuals in this class have negative attitudes toward technology and are more positive toward owning cars relative to individuals in the prior class. As

before, looking at the segment specific share of ridesharing users, it is seen that the LCCM model does an excellent job sorting between those likely to ridesharing users and those likely to be non-users, an outcome that is especially valuable for forecasting applications and shows the power of utilizing this model structure.

Table 5.10 Coefficients of downsampled NHTS LCCM with SED characteristics and predicted attitudes[a]

	Tech savvy, ridesharing-likely		**Older, pro-car, ridesharing-hesitant**	
	(53.9%)		**(46.1%)**	
	Coefficient	*p-value*	*Coefficient*	*p-value*
Outcome model				
Intercept	13.08	0.03	-1.29	0.25
Age	-0.13	0.01	-0.02	0.10
High HH income (>$100K)	2.97	0.01	0.69	0.07
HH vehicles	-2.28	0.04	0.46	0.13
Black	-3.74	0.07	0.46	0.38
Covariates				
Intercept	-0.45	0.02	0.45	0.02
Tech savvy	0.47	0.07	-0.47	0.07
Sociable	1.91	<0.001	-1.91	<0.001
Pro-car-owning	-1.27	<0.001	1.27	<0.001

[a]Note that effect coding has been used in the estimation of this model.

Table 5.11 Segment-specific shares/means of predictors and covariates for Downsampled NHTS LCCM model

Variable	**Tech savvy, ridesharing-likely (53.9%)**	**Older, pro-car, ridesharing-hesitant (46.1%)**
	Variable *means/share* per class	Variable *means/share* per class
Outcome variable		
Ridesharing usage* (binary)	0.80	0.15
Predictors		
Age †	47.42	54.41
High HH income (>$100K) *	0.48	0.33
HH vehicles †	1.86	2.10
Black *	0.28	0.19

Table 5.11 cont'd

Covariates[†]		
Tech savvy	0.47	0.21
Sociable	0.12	0.03
Pro-car-owning	-0.27	-0.07

[†] Segment-specific means
[*]Segment-specific shares (e.g., proportion of segment that can be classified as Millennials)

5.2.3 Comparison across model formulations

To close this external validation section on the attitudinal transfer variables, Table 5.12 summarizes the metrics reported for the linear regression, binary logit, and LCCM models of ridesharing usage that use both SED characteristics and predicted attitudes. As can be seen in the table, the refined LCCMs perform very similarly to the refined BLMs as far as statistical measures of fit are concerned. However, it is also seen that the LCCMs offer considerable additional behavioral insight, including more so for the less-often-chosen alternative. This may yield significant opportunities for downstream influence and analysis.

Table 5.12 Comparison across ridesharing usage models with SED and predicted attitudes (Atlanta region)

Dataset	Model	Adjusted ρ^2_{EL} ($\mathcal{L}$ (**0**) base) or adjusted R^2	Unit-weighted prediction accuracy (%)	Probability-weighted prediction accuracy (%)	Number of significant attitudes
GDOT	Linear regression	0.30	NA	NA	NA[a]
	Binary logit model	0.20	73.33	63.18	4
	LCCM	0.20	73.67	63.90	6
NHTS	Linear regression	0.19	NA	NA	NA[a]
	Binary logit model	0.51	85.77	79.23	3
	LCCM	0.50	85.32	79.03	3
NHTS Downsampled	Linear regression	0.32	NA	NA	NA[a]
	Binary logit model	0.23	75.00	65.29	2
	LCCM	0.24	76.70	67.55	3

[a]Because the linear regression models are not tuned, it is not considered appropriate to report the number of significant attitudes for these naively estimated models as they are not comparable to the significant attitudes in the other models.

5.3 Attitudinal marker variables

This last external validation effort is for the *predicted* attitudinal constructs that were transferred using the attitudinal marker variables developed in Chapter 4. As previously mentioned, marker variables were not able to be embedded into an external recipient survey like the NHTS (i.e., a proof of concept was not executed), and as such, this external validation (unlike those of the prior two sections) is not able to examine the performance of the marker variables for bringing the full set of attitudes into a recipient survey. Recall that during the construct validity internal validation detailed in Section 4.5.2, the transferred attitudinal constructs using marker statements had extremely high correlations (> 0.9) with the observed attitudinal constructs. Despite this, in Figure 5.8, it can be seen that the predicted attitudinal constructs developed using marker statements do not perform as well as the observed attitudinal constructs in improving travel behavior model lifts for the GDOT survey. Similarly, in Figure 5.9, it can be seen that the attitudinal

constructs developed using observed attitudinal indicators outperform the attitudinal constructs predicted using marker variables when attitudes alone are included as explanatory variables in the travel behavior models.

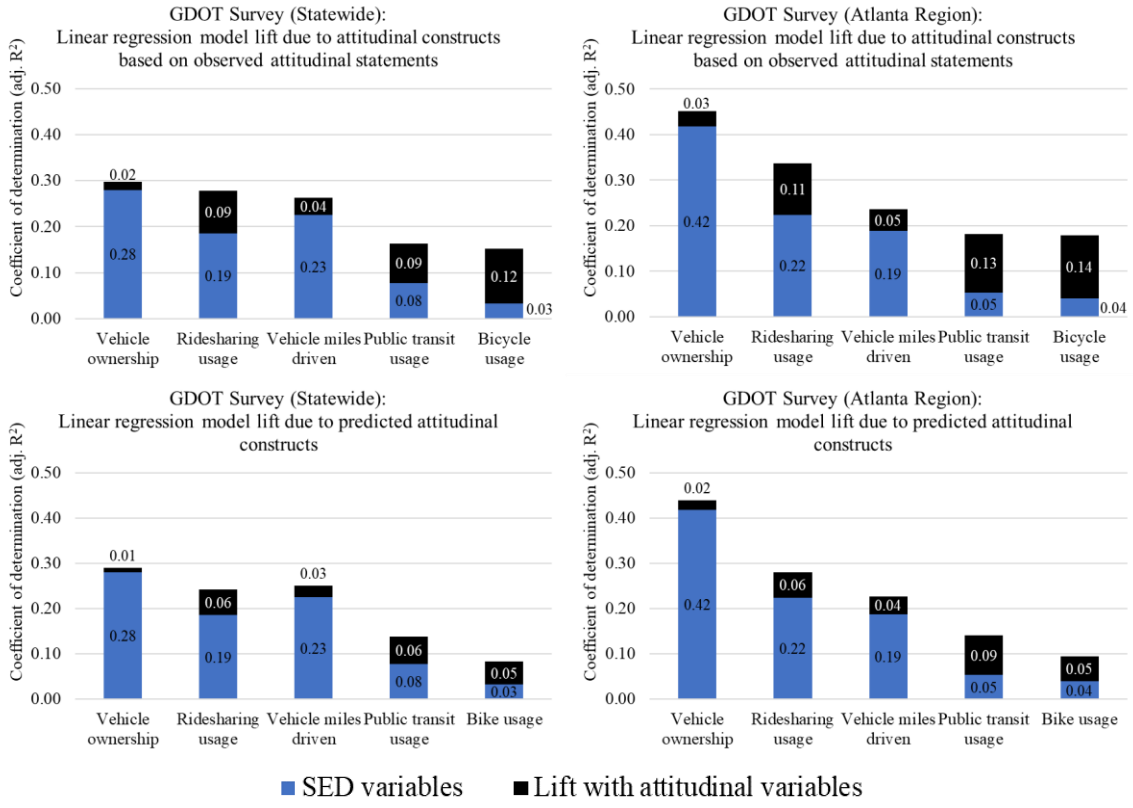

Figure 5.8. Linear regression travel behavior model lifts due to attitudinal constructs transferred using marker variables

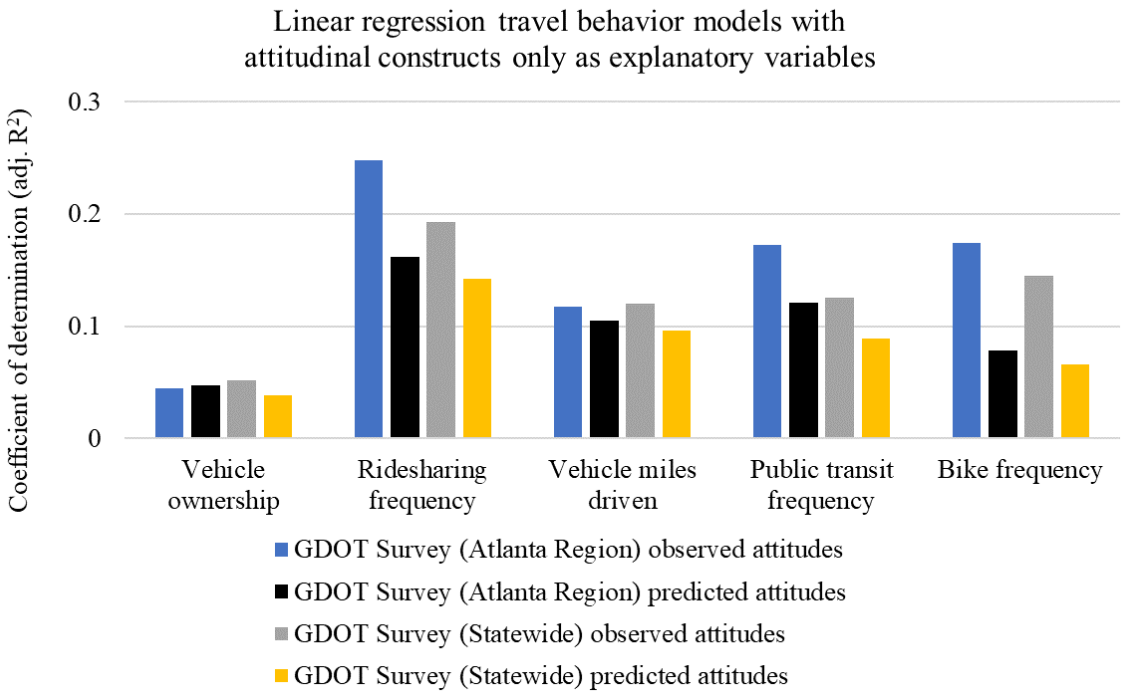

Figure 5.9. Linear regression travel behavior models with observed and marker variable predicted attitudinal constructs only as explanatory variables for the GDOT survey

5.4 Comparison across enrichment variables

Given that regression models were the only model forms executed for all three types of enrichment variables externally validated in this chapter, the adjusted R^2 values for the GDOT travel behavior models are the only performance metrics that are able to be compared relative to each other. The comparison shown here is limited to the travel behavior models executed on the Atlanta subset of the GDOT survey. In Figure 5.10, the model lifts that were obtained for the various enrichment variables throughout this chapter are displayed on the same chart. It can be seen that for vehicle ownership and vehicle miles traveled, the TM variables provide the largest model lift relative to SED-only travel behavior models. For the ridesharing, public transit, and bike usage behavioral models, it

is seen that the attitudinal constructs based on the observed attitudinal predictors (i.e., those present natively in the original survey dataset) provide the largest model lift when added to the SED variables.

Across all models, with the exception of the bicycle usage models, the attitudinal constructs that were predicted using SED, TM, and land use data provide more model lift than the attitudinal constructs that were predicted using attitudinal marker variables. This may at first be somewhat counterintuitive given that the attitudinal constructs developed using marker variables had higher correlations with the observed attitudes as shown during the internal validation process. However, this finding may be supported by the charts (Figure 5.4 and Figure 5.9) developed within the preceding sections which showed that the attitudinal constructs predicted using SED, TM, and land use outperformed the observed attitudes when attitudinal constructs were tested alone (i.e., as the sole predictors or explanatory variables) in the behavioral models, but the attitudinal constructs developed using marker variables did not do so. Thus, while this finding is difficult to fully explain, it is apparent that the transferred attitudinal constructs based on TM, SED, and land use data are bringing additional information into the model that is better at explaining travel behavior than the attitudinal constructs that are "purer" (i.e., based on observed attitudinal indicators only). Also supporting this conjecture is the fact the TM subset in the model outperforms the predicted attitudinal constructs for three of the travel behaviors models; recall that TM variables were used as inputs for the attitudinal constructs developed using SED, TM, and land use variables.

This finding shows the importance and power of external validation, as simply looking at the internal validation results would make it appear that the marker variable

process is unquestionably superior to the transfer learning process that utilizes big, external datasets. Nonetheless, it is also important to remember that the best method for enriching datasets will really depend on the intent of the application and resulting model(s). For example, if the more important use of the enrichment variables is to create interpretable models, the marker variable approach may be best. This is because the effects of predicted attitudinal constructs that are developed using an amalgamation of external, big datasets (as well as those of other explanatory variables with which they "overlap") may be more difficult to interpret as they are a result of many different variable types. However, if the analyst prioritizes model fit and model performance, the best approach may be to use the latter predicted attitudinal constructs. Thus, it can be seen that overall, using external variables directly, or using an array of methods for predicting or transferring desired variables, are all viable approaches for enriching survey datasets and improving our modeling and forecasting abilities.

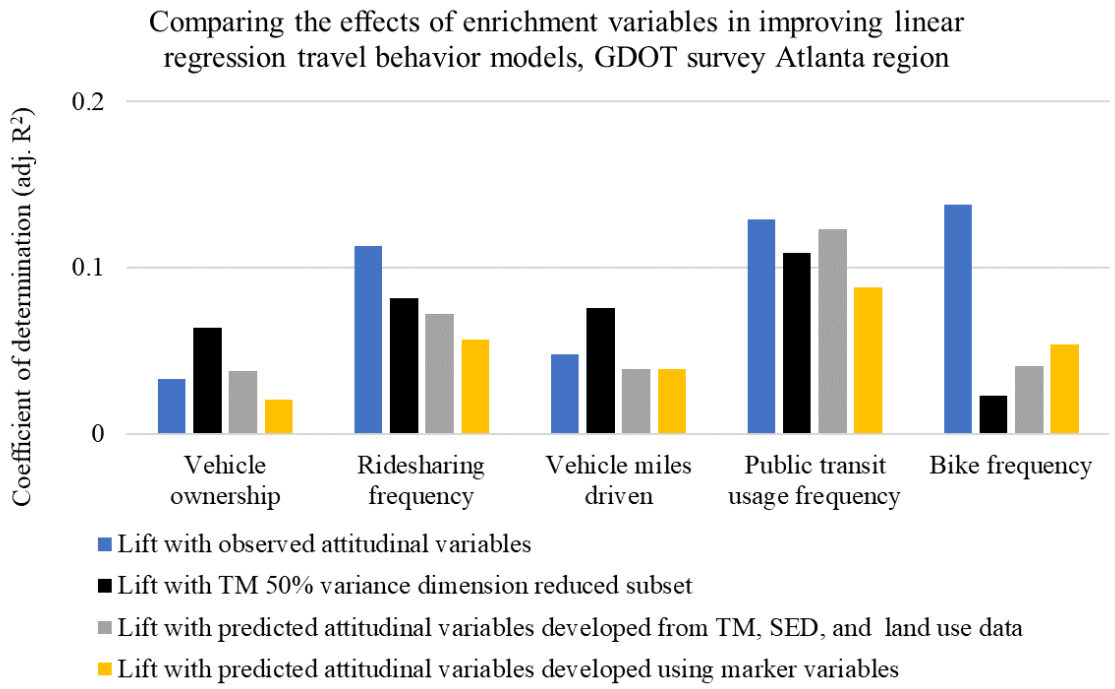

Figure 5.10. Comparison of external validation results across enrichment variables

CHAPTER 6. CONCLUSIONS

Figure 6.1 provides a visual overview of the components of this thesis, from problem definition and motivation to applications and validation. To recap, in Chapter 1, this thesis opened by identifying model constraints/limitations that may be exacerbated by transportation survey data challenges, followed by a brief discussion of possible avenues for addressing these challenges. In Chapters 2,3, and 4, three survey data enrichment methods were presented and applied. In Chapter 5, external validations of the enriched survey datasets (that resulted from the methods applied in the preceding chapters) were performed using travel behavior models. In this closing chapter (Chapter 6), a summary of the research contributions and limitations for each method and application is presented, followed by a broader discussion of future work that can expand upon the efforts made here. Lastly, a brief statement of impact is provided.

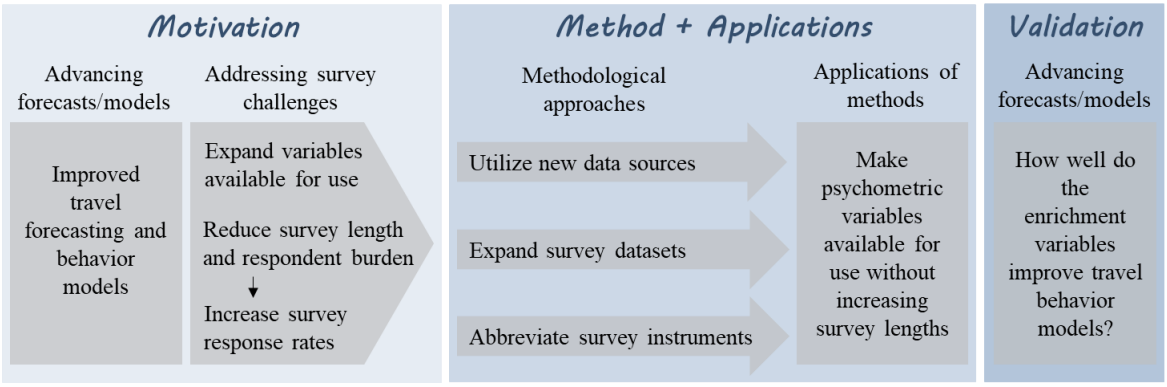

Figure 6.1. Overview of thesis components

6.1 Utilizing and integrating novel sources of data

The first methodological approach discussed in this document was that of utilizing new/novel data sources. To this end, Chapter 2 provided an example of utilizing targeted

marketing data as a potential data stream in transportation, either independently or through its deterministic linkage with survey records.

6.1.1 Contributions

A key objective of the chapter at hand (Chapter 2) is to serve as a model for analysts exploring novel data streams by providing an example of the components deemed necessary to lay the foundation for the sustainable integration of novel data sources within various fields, and specifically transportation in this case. The background components that are deemed of importance to develop for a novel data source include: a data dictionary on related terms, a detailed look at how the data is collected, from where it could be obtained, and what the benefits and disadvantages of its use might entail. It is also recommended that analysts using novel data streams develop a typology of possible applications of the data in transportation research and practice. Thereafter, depending on whether the data is to be used directly or integrated with existing data sources, a data cleaning and/or integration framework should be developed. Finally, internal validation of the data source relative to traditional data sources should be conducted. With all survey data enrichment methods, it is also important to execute external validation to assess whether the new data source provides empirical benefit to the field at hand. It is believed that only through systematic and open-sharing, can novel data sources become viable for widespread validation and usage within a field.

In this document, TM data constitutes the novel data source that is examined and for which the data sharing framework is applied. It is seen that TM data are (currently) available for most individuals in the population, are accessible for most analysts, and have

relatively low acquisition costs. In addition, the developed typology demonstrates that TM data provide a wealth of rich variables that have the potential to improve transport modeling applications and research. Internal validation procedures demonstrated that TM is able to report key SED characteristics with match rates ranging from 70% to 90% relative to traditional transportation surveys. Further, the external validation procedure illustrated the potential of TM to improve model performance for a range of travel behaviors. Thus, the application of the framework to TM data yields results that suggest wide-ranging benefits for the use of TM data in transportation.

6.1.2 Limitations

The most important limitation about this approach is the uncertainty surrounding the existence and accessibility of new data sources in the future. Changes in privacy regulations and in the method(s) of data collection used for a particular source can result in inconsistency and instability for the new data source(s). This can make it difficult for analysts to develop validation procedures that are comparable and applicable across time. Relatedly, increasingly restrictive privacy laws (for example, that allow individuals to opt in or out of data collection) can introduce additional, dynamic (i.e., ever-changing) biases within datasets. Such biases can be difficult to track since internal validation may require analysts to deterministically connect new data sources with traditional data sources, and thereafter to examine differences in variables across the datasets. As shown in Chapter 2, deterministically connecting data sources can be difficult and sometimes impossible, making it difficult to complete validation procedures.

All of these limitations listed here are applicable to TM data specifically, but are also true of a large number of passive, large data streams. Accordingly, transport professionals will increasingly have to make tradeoffs when balancing potential risks and benefits of exploring a new data source. It is hoped that discussions such as the one here provide an impetus for more civil engineers and planners to become involved with the development and implementation of privacy regulations, so as to ensure that key data sources that could significantly improve travel and urban demand models may remain accessible in some form to analysts working at city and regional levels.

6.2 Expanding survey datasets through predictive transfer

The second methodological approach discussed was the predictive transfer or imputation of variables from one survey dataset to another. In Chapter 3, this method is applied to bring attitudinal constructs from a statewide research-oriented survey (GDOT survey) into the statewide subset of a national travel survey (NHTS).

6.2.1 Contributions

The overarching objective of this probabilistic record linkage transfer learning approach is to enable analysts to transfer/predict/impute information from one survey dataset to another, using learning functions that are trained on the donor datasets and applied to the recipient datasets. At a more advanced level, this approach borrows methods from computer science to allow for the use of big datasets alongside advanced data driven models to improve the transfer of desired variables. However, it is also possible to execute the predictive transfer learning approach without these additional complexities, thus making the approach accessible for analysts across the field. The application of this method

within Chapter 3 showcased the ability to transfer complex psychometric variables across survey datasets using basic common variables, with the transfer performance improving with the addition of passive datasets like TM data.

6.2.2 Limitations

While this method is shown to have significant potential for being able to inform surveys with rich variables present in other instruments or data sources, as discussed in Section 3.6.2, there are numerous implementation challenges such as differences in spatial and temporal congruence across data sources as that can limit the applicability of the approach. Furthermore, after the method has been applied, the resultant datasets may have attributes such as undesirable distributions and substantial measurement errors that can affect downstream analyses. Furthermore, this process could lead to ambiguity in what the transferred variable actually represents, thereby resulting in the reification fallacy and/or complications in interpreting the results of downstream analyses. Information on other possible limitations of this method can be found in Section 3.6.2. With regards to the specific application shown in this thesis, there are numerous avenues for further improvement. For example, weighting can be used between datasets to improve the congruence of the distributions for key SED variables. Further, additional datasets and/or more complex ML algorithms and techniques (e.g., varied basis expansion approaches) could be applied to (potentially) improve the performance of the transfer process further. It is certainly hoped that the detailed method provided in this document serves as a foundation for future researchers, as well as myself, to further improve and advance both the method and application detailed for Chapter 3.

6.3 Abbreviating survey instruments using marker variables

The third methodological approach entailed developing condensed sets of statements (i.e., marker statements) that are representative of a larger array of observed questions/variables. These developed marker statements could be used as-is within survey instruments or alternatively could be used within the predictive transfer learning framework presented in Chapter 3 to impute the full set of constructs/statements present on the donor dataset/survey.

6.3.1 Contributions

As with the other methods detailed here, this approach is not unique, but is seldom systematically applied within transportation and/or urban planning. Thus, it is hoped that the presentation of this method alongside the appropriate terminology and visualizations would make this approach more accessible for transport researchers and practitioners. Most significantly though, the ultimate contribution and value of this approach will rely on survey designers in the field being open and willing to integrate marker statements on travel surveys that capture information not previously obtained on traditional surveys, but which could improve forecasting and behavioral modeling applications. The attitudinal marker statements developed in the application shown here represent a useful set of marker statements that could be integrated and validated on future travel surveys.

6.3.2 Limitations

All sets of marker statements developed by various teams will require external validation within varying spatial and temporal domains before the relationships between

the marker statements and the information of interest could be considered to be stable. However, this will take time and conceivably could present difficulties as varying analysts/teams will have to agree upon the marker statements selected for integration and validation across survey instruments. Thus, while this approach performed well in the limited internal and external validation tests shown in this thesis, obtaining "buy-in" in real world scenarios may be difficult. Further, if this approach utilizes the predictive transfer learning framework discussed in the prior method, as is being assumed in this thesis, then there will be additional limitations that are associated with the predictive transfer learning approach.

In line with this discussion, the major limitation of the application of the marker statement approach shown here is that it has not yet been fully externally validated for the attitudinal marker statements developed in Chapter 4. This means that the marker statements were not integrated on a second independent survey, and the subsequent relationship between the markers and the full set of information examined. As such, future work on this application will aim to first execute an independent external validation of this marker variable set. In addition, additional explorations of parameters that can be varied, for example the comparison of various structure identification approaches, would result in a more thorough application of this method.

6.4 Directions for future research

The methods presented and applied in this thesis constitute but three approaches that address various survey challenges that may limit the use of survey data in transport modeling and forecasting applications. These approaches were selected for examination in

this document because they are considered to be the most implementable and useful in bringing enriched variables into survey datasets. However, a major direction of future research that is intended to be executed next is the development of a complete typology of approaches for expanding (transportation) survey datasets. Such a typology is expected to be a valuable contribution across fields. Within transportation specifically, it could serve as a reference/guide for transport engineering and urban planning professionals interested in enriching various datasets.

Following directly from such a typology, a clear second area of future research would involve the execution of methodological application and validation of other survey enrichment methods (i.e., other methods featured on that typology) within a transport context. Examples of such methods may include:

1. survey pooling – i.e., "joining" cross sectional surveys based on similar questions; this approach has been executed by the author and other collaborators on the research team (Wang et al., under review);
2. multiple matrix sampling;
3. micro and macro-level statistical matching; and
4. multiple/simultaneous variable imputation, to name just a few that the author intends to examine next.

There are of course numerous additional survey and/or data enrichment methods that could be developed and examined.

Lastly, and from a broader perspective, based on the many challenges facing survey data, it is clear that surveys, a core research tool in many disciplines, are in need of a

widespread re-examination and subsequent plan forward to help to address errors in survey design, implementation, and analysis that are plaguing the field. One has only to look at the wildly incorrect and infamous political polls of the recent U.S. elections (2016, 2020) to get a sense of how survey errors can completely misrepresent the true values of a population. Even after the most highly respected and rigorous survey firms analyzed and corrected mistakes identified in 2016, the estimates four years later (2020) continued to be incorrect. At least one post-mortem analysis of the 2020 polls suggested that this was because the raw data continued to get worse, as weighting alone could not correct for a lack of responses from certain demographic groups (Cohn, 2020). If survey research firms with massive amounts of resources (for example, with regard to sampling reach and incentives) can still get this wrong, it forces the industry to ask, what can be done moving forward?

Unless we, as survey designers and analysts, are willing to ask ourselves the hard questions, survey data will not improve, and in fact may reach a point of being simply indefensible. This is unacceptable given that surveys are critical for building interpretable models that help explain the paradigms of how individuals make decisions and when or why they experience shifts in underlying values, traits, and attitudes. Ultimately, large, passive datasets, while extremely rich and useful, present only disjoint choices being made at specific points in time. Thus, survey data is a critical foundational data stream. However, as mentioned, over time, these data are becoming less and less representative in ways that are difficult to correct in post-analysis. Compounding this challenge, survey professionals see increasing trends of respondents simply not taking the time to provide relevant and/or thoughtful responses and thus, data quality suffers. Lastly, due to the demand for rapid

turnover and data insights, less time is spent on pretesting and ensuring that the survey questions designed are measuring what they are intended to measure (rather than actually capturing another construct).

Thus, in closing, this thesis acknowledges that the work presented here is but a tiny drop in the bucket that is needed towards improving transportation survey datasets. There are numerous avenues that need to be critically re-examined for survey data to continue to be used for the important societal decision-making processes. It is hoped that this work, and this closing plea, helps to motivate and inspire transportation researchers to take a step back from disseminating survey upon survey, and instead to take on the hard problems that may be invalidating those very results.

APPENDIX A. SUPPORTING INFORMATION FOR CHAPTER 1

Table A1. Selected sociodemographic characteristics of the survey datasets

Variable	Sample characteristics	GDOT Sample N = 3288 [a] N (%) [b]	NHTS Sample N = 5148 [a] N (%) [b]
Gender	Female	1596 (48.54)	3010 (58.47)
	Male	1678 (51.03)	2136 (41.49)
Age	18-24 years	33 (1.00)	103 (2.00)
	25-34 years	256 (7.79)	630 (12.24)
	35-44 years	330 (10.04)	737 (14.32)
	45-54 years	539 (16.39)	895 (17.39)
	55-64 years	782 (23.78)	1180 (22.92)
	65+ years	1326 (40.33)	1599 (31.06)
Tenure	Homeowner	_	3608 (70.09)
	Renter	_	1492 (28.98)
Race	Asian/Pacific Islander	58 (1.76)	80 (1.55)
	Black/African American	559 (17.00)	1219 (23.68)
	Native American	25 (0.76)	12 (0.23)
	White/Caucasian	2515 (76.49)	3537 (68.71)
Marital Status	Married	1905 (57.94)	_
	Single	1054 (32.06)	_
Dwelling Type	Stand-alone house	1905 (57.94)	_
	Apartment/condo	1054 (32.06)	_
	Mobile home	0 (0.00)	_
	Attached home/duplex/townhouse	318 (9.67)	_
Occupation	Professional managerial, or technical	1025 (31.17)	1555 (30.21)
	Sales/service	302 (9.18)	632 (12.28)
	Manufacturing, construction, maintenance, or farming	78 (2.37)	270 (5.24)
	Clerical or administrative support	121 (3.68)	306 (5.94)
Income	Less than $50,000	995 (30.26)	2437 (47.34)
	$50,000-$99,999	1151 (35.01)	1524 (29.60)
	$100,000+	1013 (30.81)	1111 (21.58)
Education	Some grade school/high school	74 (2.25)	177 (3.44)
	Completed high school or equivalent	354 (10.77)	885 (17.19)
	Some college/technical school	977 (29.71)	1557 (30.24)
	Bachelor's degree	989 (30.08)	1251 (24.3)
	Completed graduate degree(s)	887 (26.98)	1277 (24.81)
Household Size	1-person household	923 (28.07)	1890 (36.71)
	2-person household	1429 (43.46)	1923 (37.35)
	3-person household	434 (13.20)	633 (12.30)
	4- or more person household	498 (15.15)	702 (13.64)

[a] Excludes those who did not agree to be contacted again, to be consistent with the validations presented in Sections 2.6.1 and 2.6.2. An overlap sample of 1495 respondents is present in both the NHTS and GDOT survey datasets.

[b] Frequencies do not add up to 100% or the total N because of rounding errors, non-responses, or "other" categories (i.e., non-comparable categories).

APPENDIX B. SUPPORTING INFORMATION FOR CHAPTER 2

B.1. Supporting tables and figures

Table B1. Classification of TM variables (p = 5684)

Section	p (%)	Category	p (%)	Examples of subcategories
Sociodemographic	410 (7.21)	Composition	285 (5.01)	HH Structure, Age, Gender, Life stage, Background
		Education	20 (0.35)	Level, Background
		Life event	14 (0.25)	Move, Divorce, Home buyer, Relationship
		Work	21 (0.37)	Occupation, Employment status
		Housing	42 (0.74)	Length of residence, Homeowner, Codes, Density, Dwelling
		Political Indicators	28 (0.49)	Current affairs, Party membership, Political districts, Political views
Consumer-related	3453 (61.28)	Consumer Behavior	864 (15.20)	Home, Food, Automotive, Arts/Antiques, Clothing, Cause-related donations, Tobacco, Green Living, Leisure, Baby/Children, Books/magazines, Business, Channel, Classic car owner, Cost, Transaction, Home/home appliances, TV/Movie/Video, Holiday, Gift, Collectibles, Crafts, Home office/stationary, Health, Personal care, Lifestyle, General merchandise, Electronics, Novelty, Pets, Travel
		Consumer Propensity	2421 (42.59)	
		Consumer Interests	148 (2.60)	
		Consumer Attitudes	20 (0.35)	Saving, Consumerism, Shopping, Personal interest, Health, Environment
Financial	1045 (18.38)	Financial Behavior	133 (2.34)	Assets, Cash, Credit risk, Income, Insurance, Economic stability, Credit/Debit card, Mortgage, Investment, Race, Spending, Services
		Financial Propensity	880 (15.48)	Credit/Debit Card, Account, Assets, Bank, Bill, Channel, Check, Spending, Other card, Insurance, Investment, Mortgage, Offer, Service, Specification, Tax
		Attitude	32 (0.56)	Bank, Tax, Service, Investment, Economy, Financial publications
Technology[a]	205 (3.61)	Technology Behavior	18 (0.32)	Computer, Internet, Services, Other devices
		Technology Propensity	187 (3.29)	Email, Mobile phone, Mobile wallet, Service, Smart home, Channel, DVR, Social Media

Table B1 cont'd

Transport[a]	384 (6.77)	Travel Behavior	27 (0.48)	Business, Vacation, Activity, Mode, Travel purchase
		Travel Propensity	127 (2.23)	Activity, Lodging, Spending, Trip purpose, Channel, Mode, Type, Vacation
		Vehicle Behavior	42 (0.74)	Payment, Vehicle ownership, Vehicle purchase
		Vehicle Propensity	188 (3.31)	Vehicle ownership, Vehicle purchase, Vehicle rent, Loyalty, Payment, Specification, Auto club
Segmentation	184 (3.24)	Lifestyle	84 (1.48)	General segmentation[b], Health, Leisure, Shopping, Sports, Media, Food, Privacy
		Sociodemographic	20 (0.35)	Composition, Occupation, Life event
		Financial	32 (0.56)	Banking, Investment, Insurance, Affordability, Income
		Technology	29 (0.51)	Technographic[c], Technology adoption, Applications, Attitude
		Transport	19 (0.33)	Travel, Vehicle, Attitude

[a]Because technology and transportation are highly-populated categories of interest in this research domain, they are classified separately from the other consumer behavior/propensity variables.
[b]General lifestyle segmentation variables are developed based on demographic, socioeconomic, *and* consumer behaviors and are among the most well-recognized and prototypical TM variables since they capture many dimensions within one variable.
[c]The term technographic refers to general technology segmentation; in fact, the term was initially introduced in the marketing domain to characterize consumer segmentation based on attitudes, behaviors, and preferences towards technology. In addition, there is an entire lexicon devoted to technology segmentation; for example, "Mobirati" – representing the generation that cannot imagine life without mobile phones.

Table B2. Variables compared/equated between TM, NHTS, and GDOT survey data

Variable Name (Variable Type)	TM Variable Categories	NHTS Variable Categories	GDOT Variable Categories	Final Variable Categories
Gender	Male	Male	Male	Male
	Female	Female	Female	Female
	—	—	Other	Other/not applicable/missing values
	Not applicable/missing values	—	Missing values	
Age	Age relative to 2017 (2017 - birth year)	Age relative to 2017 (2017 - birth year)	Age relative to 2017 (2017 - birth year)	18-24 years
				25-34 years
				35-44 years
				45-54 years
				55-64 years
				65+ years
	Missing values	—	Missing values	Missing values
Household size [a]	Total occupants in household (number of adults + number of children)	Total occupants in household (includes non-relatives)	Total occupants in household (excludes non-relatives)	1-person household
				2-person household
				3-person household
				4- or more person household
	Missing values	Missing values	Missing values	Missing values

Table 5.2 cont'd

Income	Less than $15,000	Less than $10,000	Less than $25,000	Less than $50,000
	$15,000 to $19,999	$10,000 to $14,999		
	$20,000 to $29,999	$15,000 to $24,999	—	
	$30,000 to $39,999	$25,000 to $34,999	$25,000 to $49,999	
	$40,000 to $49,999	$35,000 to $49,999		
	$50,000 to $74,999	$50,000 to $ 74,999	$50,000 to $74,999	$50,000 to $99,999
	$75,000 to $99,999	$75,000 to $99,999	$75,000 to $99,999	
	$100,000 to $124,999	$100,000 to $124,999	$100,000 to $149,999	$100,000 or more
	Greater than $124,999	$125,000 to $149,999		
		$150,000 to $199,999	$150,000 or more	
		$200,000 or more		
	Missing values	I don't know	Missing values	Prefer not to answer/I don't know/missing values
		I prefer not to answer		
		Not ascertained		
Education	Some high school	Less than a high school graduate	Some grade school/high school	Some grade school/high school
	High school graduate	High school graduate or equivalent	Completed high school or equivalent	Completed high school or equivalent
	Some college	Some college or associates degree	Some college/technical school	Some college/technical school
	College graduate	Bachelor's degree	Bachelor's degree	Bachelor's degree
	—	—	Some graduate school	
	Graduate degree	Graduate degree or professional degree	Completed graduate degree (s)	Completed graduate degree (s)
	Missing values	Appropriate skip (age < 14)	Missing values	Not applicable/prefer not to answer/I don't know/missing values
		I don't know		
		Not ascertained		
		I prefer not to answer		
Race [b]	Asian	Asian	Asian/Pacific Islander	Asian/Pacific Islander
	Pacific Islander	Native Hawaiian or other Pacific Islander		
	Black	Black or African American	Black or African American	Black/African American
	American Indian	American Indian or Alaska native	Native American	Native American
	White	White	White/Caucasian	White/Caucasian
	Mixed (no single race/ethnic group is dominant)	Multiracial	—	Multiracial/Hispanic/other (not specified)/prefer not to answer/I don't know/missing values
	Hispanic [b]	NHTS asked Hispanic/Latino in separate question [b]	Hispanic/Latino [b]	
	Other	Some other race	Other (please specify)	
	Missing values	I don't know	Missing values	
		I prefer not to answer		

Table 5.2 cont'd

Occupation	Professional/technical	Professional managerial, or technical	Professional/technical	Professional managerial, or technical
	Self-employed - professional/technical			
	Educator			
	Financial professional			
	Legal professional			
	Medical professional			
	Administration/managerial		Manager/administrator	
	Self-employed - administration/managerial			
	Sales/service	Sales/service	Sales/marketing	Sales/service
			Services/repair	
	Self-employed – sales/service			
	Craftsman/blue collar	Manufacturing, construction, maintenance, or farming	Production/ construction	Manufacturing, construction, maintenance, or farming
	Self-employed - craftsman/blue collar			
	Farmer			
	Clerical/white collar	Clerical or administrative support	Clerical/administrative support	Clerical or administrative support
	Self-employed - clerical/white collar			
	—	—	Arts/crafts	Other/not applicable/not able to be classified/I don't know/prefer not to answer/missing values
	Other	Something else	Other	
	Self-employed - other			
	Military			
	Religious			
	Student	Appropriate skip	Not applicable	
	Self-employed - student			
	Homemaker			
	Retired			
	Self-employed - homemaker			
	Self-employed - retired			
	—	I don't know	Missing values	
	—	I prefer not to answer		
	Self-employed	Not ascertained		
Marital Status [c]	Married/inferred married	—	Married	Married
	Single/inferred single	—	Single	Single
	—	—	In a relationship	Not able to be classified/missing values
	—	—	Not single (but did not identify as married or in a relationship)	
	Missing values	—	Missing values	
Dwelling Type [c]	Single family dwelling unit	—	Stand-alone house	Stand-alone house
	Condo	—	Apartment/condo	Apartment/condo
	Apartment (5+ Units)	—		
	Mobile home	—	Mobile home	Mobile home
	2-4 United (duplex, triplex, quad)	—	Attached home/duplex/townhouse	Attached home/duplex/townhouse
	Miscellaneous residence (combo store/flat)	—	Other	Other/not able to be classified/missing values
	Timeshare	—		
	Cooperative	—		

	Missing values	—	Missing values	

Table 5.2 cont'd

Tenure/ Home ownership [c]	Homeowner	Own	—	Homeowner
	Renter	Rent	—	Renter
	—	Some other arrangement	—	Not able to be classified/I don't know/prefer not to answer/missing values
	—	I don't know	—	
	—	I prefer not to answer	—	
	Missing values	Missing values	—	

[a] NHTS data initially included non-relatives in the household size estimate; however, we adjusted this number to remove non-relatives, thus making it comparable with the GDOT data. TM did not specify whether or not non-relatives were included, but based on the variable description, which simply states that the variables encompasses the number of adults and children in a household, we have reason to believe that the TM data most likely does include non-relatives. We have accommodated this difference by instituting household tolerances in our analyses.

[b] Each of the three data sources treated the question/classification regarding "Hispanic/Latino" differently. The TM variable used treated "Hispanic/Latino" as an exclusive race (i.e., if someone was classified as "Hispanic/Latino", they could not have another race assigned to them. The NHTS asked "Hispanic/Latino" as a separate question, in keeping with the official United States Census Bureau definition that considers *race* to be "White", "Black", "Asian", "American Indian", "Pacific Islander", or another race, but considers *ethnicity* to be whether an individual is of Hispanic origin or not. As such, U.S. official documents consider race and ethnicity to be two different demographic characteristics, and typically ask them separately from each other. The GDOT survey listed "Hispanic/Latino" as another race category, but asked respondents to "*check all that apply to you"*, intending to allow "Hispanic/Latino" individuals to check another category as well. Accordingly, for the GDOT survey, we see that 82 respondents (of 3288, total) checked the "Hispanic/Latino" category, and of these 42 selected another race in addition to their identification as Hispanic/Latino. We are therefore missing race data for the remaining 40 GDOT respondents who checked Hispanic/Latino, but did not select another race. Similarly, in the TM dataset, we are missing race data for individuals classified as "Hispanic/Latino", and may similarly be missing relevant ethnicity data for individuals classified into the Census Bureau race categories. Thus, for the purposes of creating comparable race categories across these data sources, we consider all TM and GDOT respondents who are classified as "Hispanic/Latino" as "not able to be classified", and we also assign NHTS respondents who identified as "Hispanic/Latino" to the "not able to be classified" category as well.

[c] NHTS does not have an exclusive marital status or dwelling type question, and as such could not be compared to TM data. The GDOT survey does not have a tenure/home ownership question and so could not be compared to TM data.

Table B3. Variable Accuracy Rates across Overlapped Respondents before and after Processing

Variable	Match	Before Data Processing						After Data Processing					
		TM vs. GDOT N = 1495		TM vs. NHTS N = 1495		GDOT vs. NHTS N = 1495		TM vs. GDOT N = 1245		TM vs. NHTS N = 1245		GDOT vs. NHTS N = 1245	
		N	% [c]	N	%	N	%	N	%	N	%	N	%
Gender [a]	Accurate matches [b]	1366	96.40	1367	96.06	1464	98.39	1240	100	1245	100	1240	100
	Inaccurate matches [b]	51	3.60	56	3.94	24	1.61	0	0	0	0	0	0
	Not comparable [c]	78	–	72	–	7	–	5	–	0	–	5	–
Age [a]	Accurate matches	1312	94.46	1269	91.03	1411	94.76	1209	99.51	1171	95.58	1192	99.16
	Inaccurate matches	77	5.54	125	8.97	78	5.24	6	0.49	49	4.42	47	0.84
	Not comparable	106	–	101	–	6	–	30	–	25	–	6	–
Tenure [d]	Accurate matches	–	–	1330	91.72	–	–	–	–	1145	92.56	–	–
	Inaccurate matches	–	–	120	8.28	–	–	–	–	92	7.44	–	–
	Not comparable	–	–	45	–	–	–	–	–	8	–	–	–
Race	Accurate matches	1140	86.17	1129	86.65	1392	98.79	963	87.23	952	87.82	1165	99.15
	Inaccurate matches	183	13.83	174	13.35	17	1.21	141	12.77	133	12.18	10	0.85
	Not comparable	172	–	192	–	86	–	141	–	161	–	70	–
Marital status [d]	Accurate matches	981	73.48	–	–	–	–	859	74.89	–	–	–	–
	Inaccurate matches	354	26.52	–	–	–	–	288	25.11	–	–	–	–
	Not comparable	160	–	–	–	–	–	98	–	–	–	–	–
Dwelling type [d]	Accurate matches	723	60.1	–	–	–	–	616	59.52	–	–	–	–
	Inaccurate matches	480	39.9	–	–	–	–	419	40.48	–	–	–	–
	Not comparable	292	–	–	–	–	–	210	–	–	–	–	–
Occupation	Accurate matches	214	63.5	240	59.7	428	77.12	192	63.58	222	60.66	344	77.83
	Inaccurate matches	123	36.5	162	40.3	127	22.88	110	36.48	144	39.34	98	22.17
	Not comparable	1158	–	1093	–	940	–	943	–	879	–	803	–
Annual household income	Accurate matches	796	54.97	837	56.71	1170	81.19	676	56.19	706	57.45	978	81.70
	Inaccurate matches	652	45.03	639	43.29	271	18.81	527	43.81	523	42.55	219	18.30
	Not comparable	47	–	19	–	54	–	42	–	16	–	48	–
Education [e]	Accurate matches	550	45.23	555	45.53	1324	88.80	514	48.22	515	48.18	1114	89.77
	Inaccurate matches	666	54.77	664	54.47	167	11.2	552	51.78	554	51.82	127	10.23
	Not comparable	279	–	276	–	4	–	179	–	176	–	4	–

Table 5.3 cont'd

Household size [d]	Accurate matches	472	31.61	491	32.84	1251	83.79	399	32.07	406	32.61	1055	84.81
	Inaccurate matches	1021	68.39	1004	67.16	242	16.21	845	67.93	839	67.39	189	15.19
	Not comparable	2	–	0	–	2	–	1	–	0	–	1	–

[a] Gender, age (tolerance +/- 4 years), and education (tolerance: +/- 2 levels) are used in post-processing to ensure that the TM records are for the correct individuals. As such, the accuracy for these numbers in the post-processed sample are higher than would be typically expected (or unrealistically perfect, as in the case of gender). Note that even when instituting these stringent matching criteria, a large number of respondents still remain in the survey datasets. Additionally, there remain "Not comparable" cases for gender, age, and education in the post-processing sample because we retained cases for which gender/age/education are missing in either the TM or survey datasets, as these could not be definitively ruled out based on inaccurate matches.
[b] Match percentages exclude "Not comparable" segments and should be interpreted as the percentage of respondents who could be compared with an equivalent category between data sources that are accurately matched (or inaccurately matched).
[c] The "Not comparable" category includes respondents in "Other/Could not be classified/Not applicable/Prefer not to answer/Missing" categories. These categories were not separated, because they are often confounded across sources. For example, in the TM data sources, "Missing" and "Not applicable" were not distinguishable from each other, although they were distinguishable for some of the questions in the survey data sources. In Section 4.1.3 we provide general missing rates for the TM data to allow for an understanding of the rates of missingness in the TM data.
[d] NHTS did not obtain marital status and home dwelling type of survey respondents, and thus these variables could not be compared between TM and NHTS data. Similarly, GDOT did not obtain tenure, and thus this variable could not be compared between TM and GDOT survey data.
[e] When a tolerance of +/- 1 is instituted for the household size variable, the accuracy rates increase as follows: 71.40%, 70.50%, 97.12%, 71.70%, 70.60%, 97.19%.

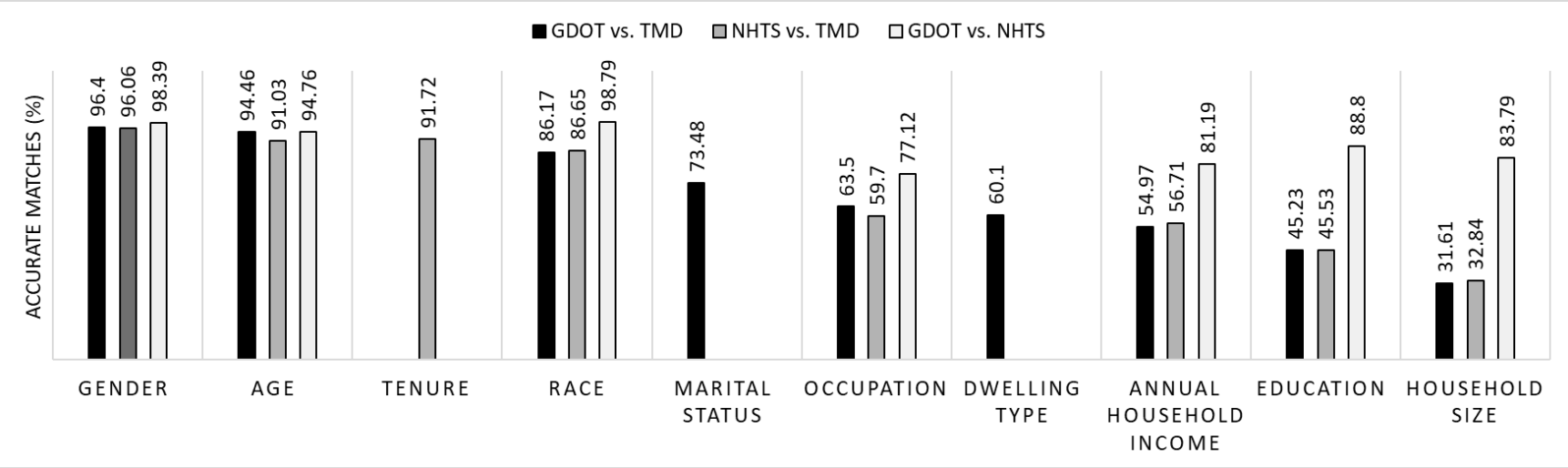

Figure B1. Variable Accuracy Rates across Overlapped Respondents before Processing

B.2. TM data integration case study

Here, we present in detail the process (Figure 5) used to integrate TM data with the GDOT survey and NHTS (see Section 4.1). We caution that the information provided *may not always* apply to all TM data providers. Given the constantly evolving and expanding nature of TM data, as well as the proliferation of new privacy laws, it is necessary for analysts to conduct their own investigations prior to making the best decision for their respective data needs and jurisdictional constraints.

B.2.1. Selection of TM data provider and service

Selecting a TM data provider depends primarily on: (1) the types of variables the analyst is hoping to acquire; and (2) the constraints of the provider regarding sample size restrictions and variable availability. There are hundreds of firms that provide TM data; however, approximately five to seven of these are mega- providers that dominate the TM data market in variable quantity and quality. These large providers are often branches of credit reporting firms such as Experian, Equifax, and TransUnion. Similar large U.S.-based TM providers that are not directly linked to credit reporting firms, but which collect/develop their own data, include Acxiom and Epsilon. Smaller TM data providers often purchase, repackage, and subsequently re-sell subsets of TM data from larger firms, and as a result data from smaller firms may not always be as regularly updated as data from larger providers. Furthermore, large TM data providers are often able to source much of their data directly, as well as to supplement their databases with valuable, high cost data from a myriad of service providers and analytics firms. Many large data providers have "small business" arms that handle queries for clients with smaller requests. Academic

researchers may often fall within these small business arms, as in our case with a sample size of N~10,000 cases. However, we recommend that academic researchers start by exploring the larger TM providers, given that even if re-directed to the small business arms, these data would most likely still be sourced from the master databases available to larger clients.

After becoming familiar with the TM firms offering the best quality data in the geographic region of interest, researchers/practitioners will need to explicitly define the type of information needed. As shown in Figure 5 of the main text, there are four primary services most typically offered by TM providers: (1) distributions of select variables by geographic region; (2) individual-name/business-name and address/ email-address lists; (3) name/email-address/phone-number appends to provided information (e.g., names appended to given addresses); and (4) data enrichment, which involves appending a wide range of selected variables to existing names and addresses. We discuss each of these services in turn. The first service listed here involves acquiring the frequencies or distributions of a particular variable, for example: gender and age, by the chosen level of geographic aggregation, whether state, county, or smaller areas like census tracts and block groups in the U.S. This service is typically used by marketing agencies who are hoping to understand the demographics of a specific area, but can also be used for aggregate level TM data validation (Kressner and Garrow, 2014).

Most transportation agencies and researchers currently use TM data primarily to purchase name and address lists in a particular jurisdiction; and there are hundreds of TM providers and other firms that provide this service. As a side note, if sampling names as well as addresses, it is important to decide whether to obtain the default name associated

with a given address in the TM database, or to request a randomly-selected (adult, usually) household member at each address – and if the latter, to double-check that that is what is actually provided. The default name is generally considered (by the TM provider) to be the "head of household", which may be adequate for many purposes. However, many survey designers obtain the names/addresses with the intention of conducting surveys on a random/representative sample of one person per household, and obtaining only heads of households would clearly preclude the study from being a random sample of the study population.

Regarding the name append service, TM firms are typically able to provide three to five selected household member names accompanied by gender and age characteristics for a given address. This service may be useful for researchers/practitioners who have addresses (and ideally SED characteristics like gender and age) for existing transport survey data, but who then wish to match this to the correct individual in order to be able to further enrich that respondent's profile. If requesting this service, it is important to inquire about the provider's match rate, i.e., the share of addresses for which names are present in the TM database, signifying that the firm has at least some information about one or more people living there.

The fourth and most relevant TM service in the context of this paper entails appending selected variables to existing names and addresses. Both large and small TM data providers typically have two options for data enrichment. For a small number of variables (e.g., 50 – 100) and limited number of respondents, there are online portals that can be used to quickly and easily append variables of interest. As the order size grows, analysts must purchase the enrichment data through "in-house" services. After engaging

with multiple data providers, our team reached the understanding that most TM firms are disinclined to provide all available variables, possibly due to proprietary concerns. However, for this project, after providing extensive justification for our use of this data, the data firm agreed to provide our team with an extensive list of variables. Our interest in obtaining almost all TM variables available was due largely to the exploratory nature of this work, and our intention to use this experience to provide guidance on the overall process as well as to identify key TM variables that may be most useful to analysts moving forward. In the future, we recommend that transportation professionals identify a smaller, more targeted subset of variables for enrichment, as this significantly simplifies the acquisition process.

Finally, when selecting a provider for data enrichment, we strongly recommend: (1) first purchasing a test set of data on a small subset of respondents; and (2) carefully and thoroughly examining the data documentation and dictionary provided. In our experience, because TM firms primarily cater to marketing clients, many of whom are interested only in reaching their target population, there tends to be a litany of variables and/or variable categories that are not accounted for, not well-defined, or simply missing from the documentation altogether. In obtaining test sets for this project, our team became aware of the fact that even providers with high quality data may not always have high quality documentation. Given that transportation professionals/researchers are often more interested than marketers in formally defining, understanding, and documenting variables, this verification is critical when it comes to selecting an appropriate provider.

B.2.2. TM data acquisition

TM providers typically require a list of names and addresses across all cases that are being submitted for data enrichment (i.e. the fourth service discussed in Section 10.2.1 and as shown in bold in Figure 5). Submitted lists are matched against names and addresses on file in the TM provider's database, and if the exact first and last name cannot be matched, variable matches degenerate into less precise matches (e.g., address and last name, address only, zip+4 code, zip code, or possibly even more aggregate matches if the attribute is not available for an exact name/address). It is therefore important to obtain and submit as complete and accurate a list of names/addresses as possible. We realize that analysts have varying amounts of name/address information available for their datasets, and thus, here we demonstrate how we dealt with the four survey data subsets in this study (see Figure 1.2 in the main text), as each had differing circumstances.

First, we examine the case in which analysts have multiple sources of name/address information available, as was the case for the GDOT_R and NHTS_Agree_R respondents. For these subsets (both of which completed the GDOT survey), there were three sources of name/address information: (1) from the original mailing list purchased for the GDOT survey (in the case of GDOT_R) or provided by GDOT (for NHTS_Agree_R); (2) from the home address question on the GDOT survey; and (3) from the final page of the GDOT survey where respondents indicated their name and address to receive a small token of appreciation. For these respondents, we developed name and address flags to cross-check names and addresses from the different sources. The flags were developed to code respondents depending on whether their self-reported names and/or addresses differed from the mailing list names and addresses to which that unique survey was delivered. For

example, if a survey addressed to person x was delivered to y address, and filled out by person z at y address, and the respondent reported her name and address accurately on the completed survey, this flag would capture the fact that the person is different, but the address to which the survey was mailed is consistent. We developed this system to help in selecting which name and address combination should be submitted for TM data augmentation.

For ~450 respondents in these subsets, duplicate cases (either two or three) were submitted for TM data augmentation due to uncertainties or differences regarding the self-reported name/address and the mailing list name/address. Duplicate cases represent a form of insurance for obtaining the best possible match rate in the data augmentation process, as once the TM variables have been appended, we can select the TM record that best matches the survey data for the individual for whom duplicate records were submitted. Thus, for analysts who have varying sources of name/address information, we recommend a similar system for choosing which record should be submitted for data enrichment. Researchers/ practitioners who have only one source of name and address information obviously would not have to engage in this checking process, as was the case for the NHTS_Agree_DNR respondents for whom we essentially had only the name/address information that they shared when agreeing to be contacted again.

The fourth subset, consisting of NHTS respondents who did not want to be contacted again (NHTS_DNAgree_DNR), required the most pre-processing, as this subset had only address information available (obtained from the trip diary data). For these individuals, we first used the TM name append service to obtain names, followed by a small-scale data enrichment to obtain gender and age, for up to five individuals living at

the submitted addresses. At this point, we investigated gender/age matches between the NHTS households and the first through fifth individuals' data obtained from the TM append services. The purpose of this process was to find the best individual match, by age and gender, between the household members in the NHTS sample and those living at the same address in the TM database. Having done this, these records can then be treated like those in the other subsets. Once the data enrichment has been completed, names can be removed from the data files as they are no longer needed.

Thus, for researchers/practitioners who have address information (which is often present in travel surveys) without names, it is possible to consider data enrichment after obtaining names linked to those addresses through a name append service. However, if no address information is available, then TM data may be integrated using a sampling approach based on selected characteristics (e.g., similar age, gender, education level, neighborhood type, etc.), a method used in synthetic population generation (see, for example, Kressner, 2017). As privacy restrictions become increasingly strict, it may be necessary to pursue flexible data enrichment approaches such as the latter option.

In this section, we provided a generalized overview of the process used to successfully enrich transportation survey data with varying amounts of name and address information. However, we acknowledge that this stage of the data enrichment process can be time consuming, and may well be a limiting factor for many users. Thus, it is recommended that analysts make appropriate plans while designing and administering their respective surveys, so that less manipulation will be necessary during post-processing.

B.2.3. TM data processing

Following data acquisition, the resulting TM dataset typically requires processing before integration with the existing datasets. The extent of these efforts is dependent on the total number of TM variables acquired, as well as on the clarity of the data documentation.

The first step entails the individual-level comparison of the TM record for each case relative to available (survey) data – this verification is intended to ensure that TM data purchased for an individual is indeed representing that individual. Analysts should first select the variables that will be pairwise compared between the TM and survey data, and subsequently establish an associated tolerance level. For example, in this paper, the three variables selected for verification were gender, age, and education level, in order of importance. After selecting the variables to be compared, we then:

1. Removed cases for which gender, age, and education level were *all* missing in the TM record, as appropriate matches between the TM data and survey data could not be established without this information.
2. Removed cases that did not match on gender.
3. Calculated age and education differences between TM and survey data for each record.
4. Retained matches with the minimum age and education differences per unique record.
5. Removed cases that fell outside of our stated tolerance levels in the following order:
    a. Removed cases that fell outside of an age tolerance of +/- four years.

b. Removed cases that fell outside of an education level tolerance of +/- two levels.

6. Conducted manual removal for unique records that still had duplicate cases remaining after the above process.

Over the course of the data matching process for this dataset, the total number of respondents was reduced by approximately 21%. Removal of cases can introduce bias; and we anticipate that retained cases are those that are more likely to have records in consumer databases (see Section 5 of the main text for examination of this).

After checking each TM record relative to the existing survey records, the TM variables should be recoded as needed. In our experience, TM databases do not always follow data conventions, and accordingly, the values are sometimes a mix of numbers and letters (e.g., a nominal variable may have "M" to represent one category and "2" to represent another, thus precluding automatic recoding). Next, variables whose share of missing values lies above a certain threshold should be either removed or imputed, depending on the analyst's final goals. Further cleaning, such as removing highly correlated variables and (near) zero variance variables, should then be conducted. Depending on the number and types of TM variables acquired, the level of cleaning and imputation required would differ. For example, since we acquired over 5000 TM variables, the imputation effort required our team to use machine learning (specifically, the Random Forest algorithm) with a supercomputer to efficiently impute across varied variable types at the same time. Traditional imputation approaches such as Expectation Maximization and Multiple Imputation are not feasible with mixed variable types across such large datasets. Additionally, for research teams that obtain large sets of TM variables, dimension

reduction or other feature/variable selection procedures may be necessary prior to using the acquired data.

APPENDIX C. SUPPORTING INFORMATION FOR CHAPTER 3

Table C1. Selected common variable sociodemographic characteristics for the transportation survey datasets

Variable	Sample characteristics	GDOT Sample N = 2699 [a] N (%) [b]	NHTS Sample N = 4581 [a] N (%) [b]
Gender	Female	1272 (47.12)	2707 (59.09)
	Male	1427 (52.87)	1874 (40.91)
Generation	18-34 years	197 (7.30)	480 (10.48)
	35-44 years	253 (9.37)	621 (13.56)
	45-64 years	1098 (40.68)	1931 (42.15)
	65+ years	1151 (42.65)	1549 (33.81)
Race	Asian/Pacific Islander	47 (1.74)	101 (2.20)
	Black/African American	471 (17.45)	1048 (22.88)
	Native American	55 (2.04)	78 (1.70)
	White/Caucasian	2175 (80.59)	3431 (74.90)
Occupation	Professional managerial, or technical	966 (35.79)	1424 (31.08)
	Sales/service	289 (10.71)	549 (11.98)
	Manufacturing, construction, maintenance, or farming	62 (2.30)	278 (6.07)
	Clerical or administrative support	112 (4.15)	271 (5.92)
Worker	Worker	1460 (54.09)	2526 (55.14)
Income	Less than $25,000	324 (12.00)	983 (21.46)
	$25,000 to $49,999	548 (20.30)	1059 (23.12)
	$50,000 to $74,999	552 (20.45)	841 (18.36)
	$75,000 to $99,999	423 (15.67)	594 (12.97)
	$100,000 to $149,999	497 (18.41)	665 (14.52)
	$150,000 or more	355 (13.15)	439 (9.58)
Education	Some grade school/high school	48 (1.78)	131 (2.86)
	Completed high school or equivalent	295 (10.93)	913 (19.93)
	Some college/technical school	833 (30.86)	1465 (31.98)
	Bachelor's degree	839 (31.09)	1098 (23.97)
	Completed graduate degree(s)	684 (25.34)	974 (21.26)
Driver	Driver	2673 (99.04)	4361 (95.20)
Number of household drivers	No drivers	12	136
	1 driver	862	1683
	2 drivers	1412	2302
	3 drivers	269	361
	4+ drivers	144	99
Household Size	1-person household	784 (29.05)	1385 (30.23)
	2-person household	1184 (43.87)	1938 (42.31)
	3-person household	333 (12.34)	603 (13.16)
	4- or more person household	398 (14.75)	220 (4.80)

Table C2. Attitudinal indicators and latent constructs for 15- factor EFA and CFA solution

Factor	Statement	EFA Loading	CFA Loading[a]
Non-car alternatives	s. I like the idea of walking as a means of travel for me.	0.730	0.805
	ae. I like the idea of bicycling as a means of travel for me.	0.727	0.680
	c. I like the idea of public transit as a means of travel for me.	0.350	0.577
Tech-savvy	g. Learning how to use new technologies is often frustrating for me.	-0.938	-0.778
	af. I am confident in my ability to use modern technologies.	0.835	0.974
Commute benefit	y. My commute is a useful transition between home and work (or school).	0.693	0.732
	q. My travel to/from work (or school) is usually pleasant.	0.610	0.641
	as. I wish I could instantly be at work (or school) – the trip itself is a waste of time.	-0.421	-0.448
Modern urbanite	l. I like the idea of having stores, restaurants, and offices mixed among the homes in my neighborhood.	0.432	0.240
	k. My phone is so important to me, it's almost part of my body.	0.398	0.873
Work-oriented	d. At this stage of my life, having fun is more important to me than working hard.	-0.475	-0.343
	u. I'm too busy to have as much leisure time as I'd like.	0.675	0.877
Materialistic	ah. I usually go for the basic (“no-frills”) option rather than paying more money for extras.	-0.598	-0.620
	n. The functionality of a car is more important to me than the status of its brand.	-0.451	-0.455
	z. I would/do enjoy having a lot of luxury things.	0.417	0.508
	aq. I like to wait a while rather than being first to buy new products.	-0.364	-0.397
	b. I prefer to minimize the amount of things I own.	-0.344	-0.442
Polychronic	ag. I prefer to do one thing at a time.	-0.919	-0.898
	e. I like to juggle two or more activities at the same time.	0.725	0.662
Pro-environmental	v. Cost or convenience takes priority over environmental impacts (e.g. pollution) when I make my daily choices.	-0.941	0.538
	ar. I am committed to an environmentally-friendly lifestyle.	0.550	-0.918
Family/friends oriented*	p. Family/friends play a big role in how I schedule my time.	-0.602	0.520
	w. It’s okay to give up a lot of time with family and friends to achieve other worthy goals.	0.467	-0.565
Pro-suburban	aa. I prefer to live in a spacious home, even if it’s farther from public transportation or many places I go to.	0.651	0.849
	f. I see myself living long-term in a suburban or rural setting.	0.362	0.310
	z. I would/do enjoy having a lot of luxury things.	0.439	0.530
Waiting-tolerant*	al. Having to wait is an annoying waste of time.	0.958	-0.861
	h. Having to wait can be a useful pause in a busy day.	-0.526	0.564
Travel-liking*	ac. I generally enjoy the act of traveling itself.	-0.716	0.264
	a. I like exploring new places.	-0.563	0.394
Sociable*	x. I consider myself to be a sociable person.	-0.687	-0.490
	o. I’m uncomfortable being around people I don’t know.	0.462	0.359
Pro-car-owning	t. I definitely want to own a car.	0.882	0.901
	j. I am fine with not owning a car, as long as I can use/rent one any time I need it.	-0.599	-0.655
	ak. I like the idea of driving as a means of travel for me.	0.460	0.589

Table C2 cont'd

Pro-exercise	ao. The importance of exercise is overrated.	0.756	0.685
	m. I am committed to exercising regularly.	-0.702	-0.813

[a] Not all factors used in the CFA are shown here. For simplicity, we include only the loadings for the statements that overlap with the EFA construct statements.
*The loadings on these statements must be reversed during interpretation.They have been reversed as needed in all model results shown.

Table C3. Attitudinal indicators and latent constructs for six-factor solution

Factor	Statement	EFA Loading	CFA Loading[a]
Non-car alternatives	s. I like the idea of walking as a means of travel for me.	0.560	0.752
	ae. I like the idea of bicycling as a means of travel for me.	0.491	0.636
	c. I like the idea of public transit as a means of travel for me.	0.565	0.682
	m. I am committed to exercising regularly.	0.431	0.459
	a. I like exploring new places.	0.388	0.345
	ar. I am committed to an environmentally-friendly lifestyle.	0.340	0.364
	l. I like the idea of having stores, restaurants, and offices mixed among the homes in my neighborhood.	0.339	0.432
	ac. I generally enjoy the act of traveling itself.	0.302	--
Tech-savvy	g. Learning how to use new technologies is often frustrating for me.	-0.963	-0.836
	af. I am confident in my ability to use modern technologies.	0.736	0.904
Commute and wait tolerant	as. I wish I could instantly be at work (or school) – the trip itself is a waste of time.	-0.613	0.536
	al. Having to wait is an annoying waste of time.	-0.569	0.614
	y. My commute is a useful transition between home and work (or school).	0.495	-0.584
	h. Having to wait can be a useful pause in a busy day.	0.478	-0.562
	q. My travel to/from work (or school) is usually pleasant.	0.432	-0.537
Polychronic	ag. I prefer to do one thing at a time.	1.058	0.936
	e. I like to juggle two or more activities at the same time.	-0.557	-0.640
Materialistic	z. I would/do enjoy having a lot of luxury things.	-0.670	0.687
	ah. I usually go for the basic ("no-frills") option rather than paying more money for extras.	0.483	-0.507
	aa. I prefer to live in a spacious home, even if it's farther from public transportation or many places I go to.	-0.391	0.530
	n. The functionality of a car is more important to me than the status of its brand.	0.388	-0.380
	b. I prefer to minimize the amount of things I own.	0.366	-0.458
	aq. I like to wait a while rather than being first to buy new products.	0.316	-0.325
Pro-car owning	ak. I like the idea of driving as a means of travel for me.	0.621	0.656
	t. I definitely want to own a car.	0.599	0.816
	aj. As a general principle, I'd rather own things myself than rent or borrow them from someone else.	0.479	0.569
	j. I am fine with not owning a car, as long as I can use/rent one any time I need it.	-0.446	-0.663
	f. I see myself living long-term in a suburban or rural setting.	0.362	0.391

[a] Not all factors used in the CFA are shown here. For simplicity, we include only the loadings for the statements that overlap with the EFA construct statements.

Table C4. Native common variables in GDOT and NHTS surveys

Variable Name	NHTS Variable Categories	GDOT Variable Categories	Final Variable Categories
Gender	Male	Male	Male
	Female	Female	Female
Age	Age relative to 2017 (2017 - birth year)	Age relative to 2017 (2017 - birth year)	Age relative to 2017 (2017 - birth year)
Household size [a]	Total occupants in household (includes non-relatives)	Total occupants in household (excludes non-relatives)	Total occupants in household (excludes non-relatives)
Household income	Less than $10,000	Less than $25,000	Less than $25,000
	$10,000 to $14,999		
	$15,000 to $24,999	—	
	$25,000 to $34,999	$25,000 to $49,999	$25,000 to $49,999
	$35,000 to $49,999		
	$50,000 to $ 74,999	$50,000 to $74,999	$50,000 to $74,999
	$75,000 to $99,999	$75,000 to $99,999	$75,000 to $99,999
	$100,000 to $124,999	$100,000 to $149,999	$100,000 to $149,999
	$125,000 to $149,999		
	$150,000 to $199,999	$150,000 or more	$150,000 or more
	$200,000 or more		
Education	Less than a high school graduate	Some grade school/high school	Some grade school/high school
	High school graduate or equivalent	Completed high school or equivalent	Completed high school or GED
	Some college or associates degree	Some college/technical school	Some college/technical school
	Bachelor's degree	Bachelor's degree	Bachelor’s degree
	—	Some graduate school	
	Graduate degree or professional degree	Completed graduate degree (s)	Completed graduate degree (s)
	Appropriate skip (age < 14)	Missing values	Not applicable/missing values
	I don't know		
	Not ascertained		
	I prefer not to answer		
Race [b]	Asian	Asian/Pacific Islander	Asian/Pacific Islander
	Native Hawaiian or other Pacific Islander		
	Black or African American	Black or African American	Black/African American
	American Indian or Alaska native	Native American	Native American
	White	White/Caucasian	White/Caucasian
	Multiracial	More than one category selected	Multiracial
	NHTS asked Hispanic/Latino in separate question [b]	Hispanic/Latino [b]	Hispanic/Latino [b]
	Some other race	Other (please specify)	Other
	I don’t know	Missing values	Missing
	I prefer not to answer		
Worker[c]	Worker	Worker	Worker

Table C4 cont'd

Work full-time for pay[c]	Work full- time for pay	Work full- time for pay	Work full- time for pay
Work part-time for pay[c]	Work part-time for pay	Work part-time for pay	Work part-time for pay
Work two jobs[c]	Work more than one job	I have two or more paying jobs.	Work two jobs
Homemaker[c]	A homemaker	I am a homemaker/caregiver	Homemaker
Student[c]	Going to school	Full time/part time student	Student
Retired[c]	Retired	I am retired	Retired
Other work	Something else	I do unpaid work	Other work
		Other	
Medical condition	Given up driving altogether	Physical conditions or anxieties that absolutely prevents driving during the day	Respondent has a condition, handicap, anxieties that prevent or limit them from traveling outside the home
		Physical conditions or anxieties that absolutely prevents driving at night	
		Physical conditions or anxieties that absolutely prevents driving on the freeway	
	—	Physical conditions or anxieties that absolutely prevents taking public transit	
	—	Physical conditions or anxieties that absolutely prevents walking	
	—	Physical conditions or anxieties that absolutely prevents riding a bicycle	
	—	Physical conditions or anxieties that limits driving during the day	
	Limited driving to daytime	Physical conditions or anxieties that limits driving at night	
	—	Physical conditions or anxieties that limits driving on the freeway	
	Used the bus or subway less frequently	Physical conditions or anxieties that limits taking public transit	
	—	Physical conditions or anxieties that limits walking	
	—	Physical conditions or anxieties that limits riding a bicycle	
	Reduced day-to-day travel	—	
	Asked others for rides	—	
	Used special transportation services such as Dial-A-Ride	—	
	Used a reduced fare taxi	—	
Driver	Driver? Yes	Age you received your license	Yes, I am a driver
	Driver? No	I don't have a license	No, I am not a driver
HH driver	Number of related people in HH who are drivers	Number of related people in HH who hold a license	Number of relatives in HH who are drivers
Household age groups[c]	() persons under 6	() persons under 6	() persons under 6
	() persons 6-12	() persons 6-12	() persons 6-12
	() persons 15-17	() persons 15-17	() persons 15-17
	() persons 18-26	() persons 18-26	() persons 18-26
	() persons 27-34	() persons 27-34	() persons 27-34
	() persons 35-50	() persons 35-50	() persons 35-50
	() persons 51-65	() persons 51-65	() persons 51-65
	() persons over 65	() persons over 65	() persons over 65

Table C4 cont'd

[a] NHTS data initially included non-relatives in the household size estimate; however, we adjusted this number to remove non-relatives, thus making it comparable with the GDOT data.
[b] The NHTS asked "Hispanic/Latino" as a separate question, in keeping with the official United States Census Bureau definition that considers *race* to be "White", "Black", "Asian", "American Indian", "Pacific Islander", or another race, but considers *ethnicity* to be whether an individual is of Hispanic origin or not. As such, U.S. official documents consider race and ethnicity to be two different demographic characteristics, and typically ask them separately of each other. The GDOT survey listed "Hispanic/Latino" as another race category, but asked respondents to "*check all that apply to you"*, intending to allow "Hispanic/Latino" individuals to check another category as well. Accordingly, for the GDOT survey, we see that 64 respondents (of 2699, total) checked the "Hispanic/Latino" category, and of these 27 selected another race in addition to their identification as Hispanic/Latino. The original dataset was therefore missing race data for the remaining 37 GDOT respondents who checked Hispanic/Latino, but did not select another race. In the initial processing of the dataset, race data for these 37 respondents was imputed.
[c]These variables did not need adjustment of categories between GDOT and NHTS datasets.

APPENDIX D. SUPPORTING INFORMATION FOR CHAPTER 5

Table D1. Travel behavior variables across GDOT and NHTS surveys

Variable name	NHTS question (response type/categories)	GDOT question[a] (response type/categories)	Final response categories for harmonized variables
Vehicle ownership	How many cars (including light truck, minivan, SUV, and motorcycle) does your household have? (continuous variable)	How many vehicles are owned, leased, or available for regular use by the people who currently live in your household? Include motorcycles, mopeds and RVs. (continuous variable)	No harmonization needed; continuous variable
Ridesharing usage	In the past 30 days, how many times have you purchased a ride with a smartphone rideshare app (e.g., Uber, Lyft, Sidecar, etc.) (continuous variable)	Please indicate how often you typically use on-demand ride services or shared on-demand ride services. (Answer options: never used/no longer use; less than once a month; 1-3 times a month; 1-2 times a week; 3 or more time a week)	Never used/no longer use Less than once a month 1-3 times a month 1-2 times a week 3 or more times a week
Carsharing	In the past 30 days, how many time did you use a carsharing service where a car can be rented by the hour (e.g., Zipcar or Car2Go? (continuous variable)	Please indicate how often you typically use carsharing. (Answer options: never used/no longer use; less than once a month; 1-3 times a month; 1-2 times a week; 3 or more time a week)	Never used/no longer use Less than once a month 1-3 times a month 1-2 times a week 3 or more times a week
Public transit	In the past 30 days, about how many days have you used public transportation such as buses, subways, streetcars, or commuter trains? (continuous variable)	Please indicate how often you typically make local (i.e., not overnight) trips using bus or train. (Answer options: never; less than once per month; 1-3 times a month; 1-2 times a week; 3-4 times a week; 5 or more (5-7) times a week)	Never used/no longer use Less than once a month 1-3 times a month 1-2 times a week 3 -4 times a week 5 or more (5-7) times a week
Bicycle	In the past 7 days, how many times did you ride a bicycle outside including bicycling to exercise, or to go somewhere (e.g., bike to a friend's house, bike around the neighborhood, bike to the store, etc.)? (continuous variable)	Please indicate how often you typically make local (i.e., not overnight) trips using bicycle. (Answer options: never used/no longer use; less than once per month; 1-3 times a month; 1-2 times a week; 3-4 times a week; 5 or more (5-7) times a week)	Never used/no longer use Less than once a month 1-3 times a month 1-2 times a week 3 -4 times a week 5 or more (5-7) times a week

Table D1 cont'd			
Vehicle miles driven	Please provide your best guess as to how many miles you personally drove during the past 12 months in all motorized vehicles. Include all miles from work vehicles, rental cars, and any other vehicles not owned by your household. (continuous variable)	Now considering your travel for all purposes, how many miles do you personally drive in a typical week? If you are a professional driver (e.g., bus, truck, taxi, or Uber/Lyft driver), please do not include the miles you cover as part of your job. (continuous variable)	Not able to be harmonized across surveys due to reporting differences for work-related miles driven; reported only for GDOT survey in this thesis

[a]Question formulations shown here may not look exactly like this on the survey instrument due to graphical limitations that prevent exact reproduction. However, the essence and information that the question is capturing is accurately represented here.

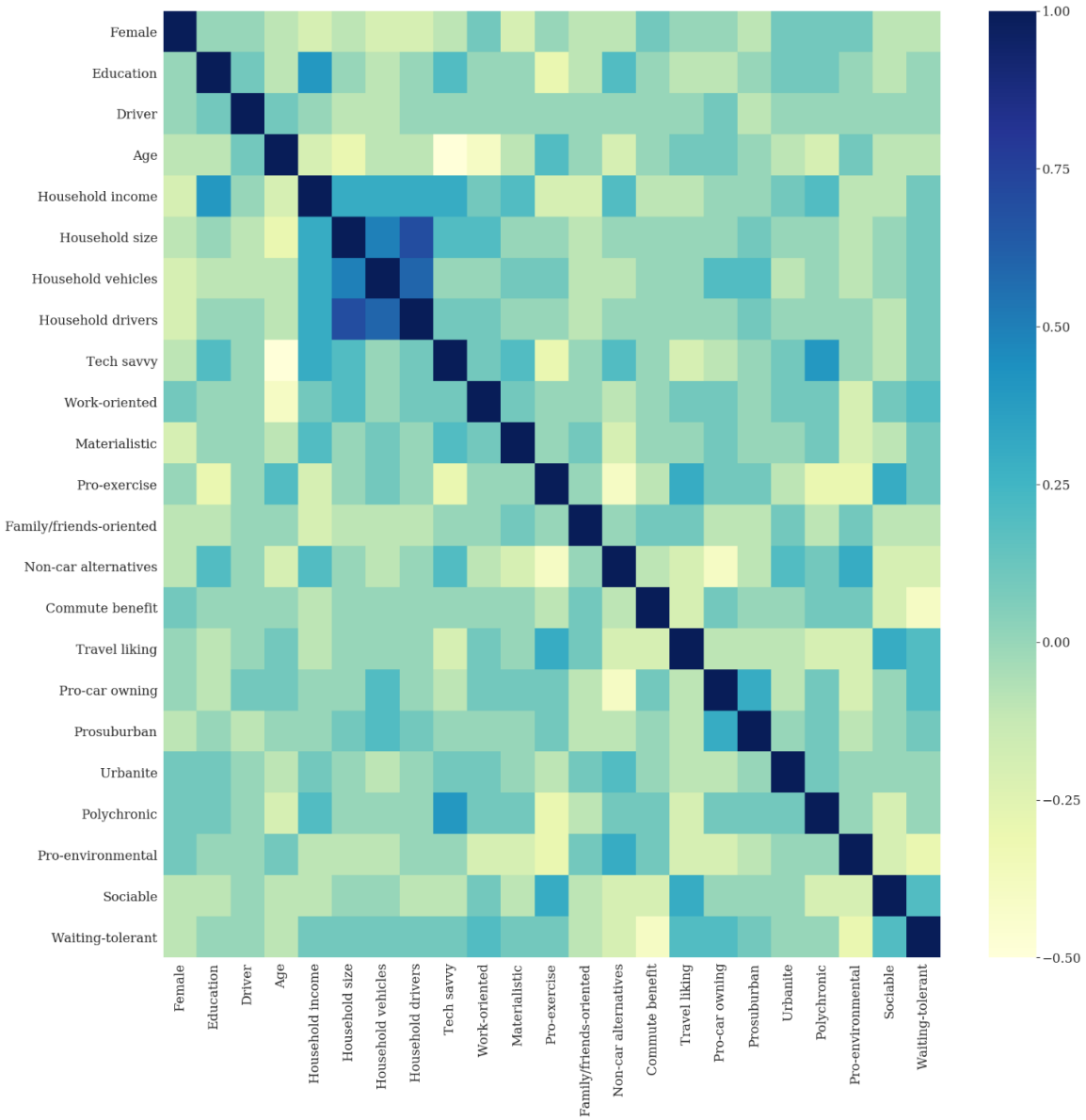

Figure D1. Correlations for GDOT SED characteristics and observed attitudinal constructs

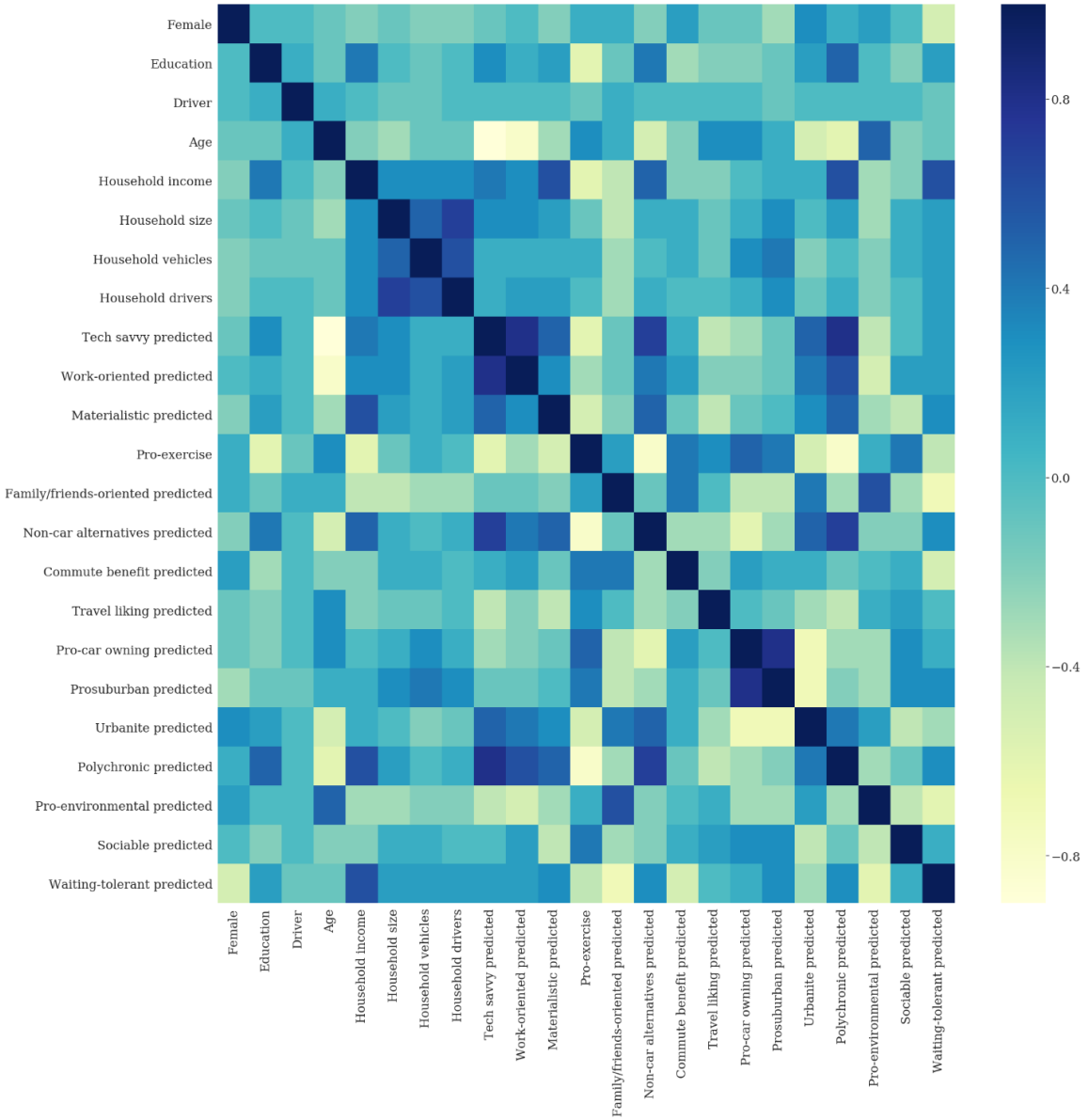

Figure D2. Correlations for GDOT SED characteristics and predicted attitudinal constructs

REFERENCES

Aalto, M., Alho, H., Halme, J. T., & Seppä, K. (2009). AUDIT and its abbreviated versions in detecting heavy and binge drinking in a general population survey. *Drug and Alcohol Dependence, 103*(1), 25-29. doi: 10.1016/j.drugalcdep.2009.02.013

Acxiom. (2020). General Data Protection Regulation Privacy Notice. Retrieved January 22, 2020 from: https://www.acxiom.com/about-us/privacy/gdpr/

Adriaan, H., & Jacco, D. (2009). Nonresponse in the recruitment of an internet panel based on probability sampling. *Survey Research Methods*, *3*(2), 59-72. doi: 10.18148/srm/2009.v3i2.1551

Alemi, F., Circella, G., Mokhtarian, P., & Handy, S. (2019). What drives the use of ridehailing in California? Ordered probit models of the usage frequency of Uber and Lyft. *Transportation Research Part C: Emerging Technologies, 102*, 233-248. doi: 10.1016/j.trc.2018.12.016

Amarov, B., & Rendtel, U. (2013). The recruitment of the access panel of German official statistics from a large survey in 2006: Empirical results and methodological aspects. *Survey Research Methods, 7*(2)**,** 103-114. doi: 10.18148/srm/2013.v7i2.5069

Anable, J., & Wright, S. (2013). *Golden Questions and Social Marketing Guidance Report*. Report, Centre for Transport Research, available from https://abdn.pure.elsevier.com/en/publications/golden-questions-and-social-marketing-guidance-report

Bain, R. (2009). Error and optimism bias in toll road traffic forecasts. *Transportation, 36*, 469-482. doi:10.1007/s11116-009-9199-7

Basarkod, G., Sahdra, B., & Ciarrochi, J. (2018). Body Image–Acceptance and Action Questionnaire–5: An Abbreviation Using Genetic Algorithms. *Behavior Therapy, 49*(3), 388-402. doi: 10.1016/j.beth.2017.09.006

Beckman, R. J., Baggerly, K. A., & McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice, 30*(6), 415-429. https://doi.org/10.1016/0965-8564(96)00004-3

Bhandari, A. (2020). *Feature Scaling for Machine Learning: Understanding the Difference Between Normalization vs. Standardization*. Retrieved January 21, 2021 from: https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/

Binder, S., Macfarlane, G. S., Garrow, L. A., & Bierlaire, M. (2014). Associations among household characteristics, vehicle characteristics and emissions failures: An application of targeted marketing data. *Transportation Research Part A: Policy and Practice, 59*, 122-133. doi: 10.1016/j.tra.2013.11.005

Birkin, M., Morris, M., Birkin, T., & Lovelace, R. (2017). Using census data in microsimulation. In J. Stillwell (Ed.), *Census Users Handbook 2011*. Ashgate, London.

Birkin, M. (2019). Spatial data analytics of mobility with consumer data. *Journal of Transport Geography, 76*, 245-253. doi: 10.1016/j.jtrangeo.2018.04.012

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York, NY: Springer Science+Business Media, LLC.

Cahan, E. M., Hernandex-Boussard, T., Thadaney-Israni, S., & Rubin, D. L. (2019). Putting the data before the algorithm in big data addressing personalized healthcare. *npj Digital Medicine*, *2*(78). doi: 10.1038/s41746-019-0157-2

Cain, K. L., Gavand, K. A., Conway, T. L., Geremia, C. M., Millstein, R. A., Frank, L. D., . . . Sallis, J. F. (2017). Developing and validating an abbreviated version of the Microscale Audit for Pedestrian Streetscapes (MAPS-Abbreviated). *Journal of Transport & Health, 5*, 84-96. doi: 10.1016/j.jth.2017.05.004

Center for Neighborhood Technology. (2019). AllTransit™. https://alltransit.cnt.org

Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies, 68*, 285-299. doi: 10.1016/j.trc.2016.04.005

Cheng, L., Chen, X., De Vos, J., Lai, X., & Witlox, F. (2019). Applying a random forest method approach to model travel mode choice behavior. *Travel Behaviour and Society, 14*, 1-10. doi: 10.1016/j.tbs.2018.09.002

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, N.J.: L. Erlbaum Associates.

Cohn, N. (2020). What went wrong with polling? Some early theories. Retrieved January 22, 2021from: https://www.nytimes.com/2020/11/10/upshot/polls-what-went-wrong.html

Connelly, R., Playford, C. J., Gayle, V., & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research, 59*, 1-12. doi: 10.1016/j.ssresearch.2016.04.015

Couper, M. P., Kapteyn, A., Schonlau, M., & Winter, J. (2007). Noncoverage and nonresponse in an internet survey. *Social Science Research*, *36*(1), 131-148. doi: 10.1016/j.ssresearch.2005.10.002

The Data Detective. (2020). *The 80/20 Split Intuition and an Alternative Split Method*. Retrieved December 18, 2020 from: https://towardsdatascience.com/finally-why-we-use-an-80-20-split-for-training-and-test-data-plus-an-alternative-method-oh-yes-edc77e96295d

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method.* New York: John Wiley & Sons, Inc.

D'Orazio, M., Di Zio, M., and Scanu, M. (2006*). Statistical Matching: Theory and Practice*. Chichester, England: John Wiley & Sons, Inc.

D'Orazio, M. (2017). Statistical Matching and Imputation of Survey Data with StatMatch.

Domarchi, C., Tudela, A., & González, A. (2008). Effect of attitudes, habit and affective appraisal on mode choice: An application to university workers. *Transportation, 35*, 585-599. doi:10.1007/s11116-008-9168-6

Eisenbarth, H., Lilienfeld, S. O., & Yarkoni, T. (2015). Using a Genetic Algorithm to Abbreviate the Psychopathic Personality Inventory-Revised (PPI-R). *Psychol Assess, 27*(1), 194-202. doi:10.1037/pas0000032

Eisenmann, C., & Kuhnimhof, T. (2018). Some pay much but many don't: Vehicle TCO imputation in travel surveys. *Transportation Research Procedia, 32*, 421-435. doi: 10.1016/j.trpro.2018.10.056

Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big data consumer analytics and the transformation of marketing. *Journal of Business Research, 69*(2), 897-904. doi: 10.1016/j.jbusres.2015.07.001

Federal Highway Administration. (2018). National Household Travel Survey. Retrieved July 23, 2019 from https://nhts.ornl.gov

Feng, J., Wang, Y., Peng, J., Sun, M., Zeng, J., & Jiang, H. (2019). Comparison between logistic regression and machine learning algorithms on survival prediction of traumatic brain injuries. *Journal of Critical Care*, 54, 110-116. doi: 10.1016/j.jcrc.2019.08.010

Google Developers. (2020). Data preparation and feature engineering for machine learning: imbalanced data [Online course]. Retrieved January 21, 2021 from:

https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data

Goyder, J., Boyer, L., & Martinelli, G. (2006). Integrating Exchange and Heuristic Theories of Survey Nonresponse. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique, 92*(1), 28-44. doi:10.1177/075910630609200104

Handy, S., Cao, X., & Mokhtarian, P. (2005). Correlation or causality between the built environment and travel behavior? Evidence from Northern California. *Transportation Research Part D: Transport and Environment, 10*(6), 427-444. doi: 10.1016/j.trd.2005.05.002

Hartgen, D. T. (2013). Hubris or humility? Accuracy issues for the next 50 years of travel demand modeling. *Transportation, 40*(6), 1133-1157. doi:10.1007/s11116-013-9497-y

Hastie, T., Tibshirani, R., and Friedman, J. (2016). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer Science+Business Media.

He, S. Y., Miller, E. J., & Scott, D. M. (2018). Big data and travel behaviour. *Travel Behaviour and Society, 11*, 119-120. doi: 10.1016/j.tbs.2017.12.003

Hope, A. C. A. (1968). A Simplified Monte Carlo Significance Test Procedure. *Journal of the Royal Statistical Society. Series B (Methodological), 30*(3), 582-598. doi: 10.1111/j.2517-6161.1968.tb00759.x

Hössinger, R., Aschauer, F., Jara-Díaz, S., Jokubauskaite, S., Schmid, B., Peer, S., . . . Gerike, R. (2020). A joint time-assignment and expenditure-allocation model: value of leisure and value of time assigned to travel for specific population segments. *Transportation, 47*(3), 1439-1475. doi:10.1007/s11116-019-10022-w

Kelly, F., & Doriot, P. (2017). *Using data mining and machine learning to shorten and improve surveys.* Retrieved January 22, 2021 from https://www.quirks.com/articles/using-data-mining-and-machine-learning-to-shorten-and-improve-surveys

Kerckhoffs, J., Hoek, G., Portengen, L., Brunekreef, B., & Vermeulen, R. C. H. (2019). Performance of Prediction Algorithms for Modeling Outdoor Air Pollution Spatial Surfaces. *Environmental Science and Technology*, 53(3), 1413-1421. doi:10.1021/acs.est.8b06038

Khan, S. M., Ngo, L. B., Morris, E. A., Dey, K., & Zhou, Y. (2017). Social media data in transportation. In M. Chowdhury, A. Apon, & K. Dey (Eds.), *Data Analytics for Intelligent Transportation Systems* (pp. 263-281): Elsevier.

Kitamura, R. (2009). Life-style and travel demand. *Transportation, 36*(6), 679-710. doi: 10.1007/s11116-009-9244-6

Kim, S. H., & Mokhtarian, P. L. (2018). Taste heterogeneity as an alternative form of endogeneity bias: Investigating the attitude-moderated effects of built environment and socio-demographics on vehicle ownership using latent class modeling. *Transportation Research Part A: Policy and Practice, 116*, 130-150. doi: 10.1016/j.tra.2018.05.020

Kim, S., Mokhtarian, P. L., & Circella, G. (2019). *The Impact of Emerging Technologies and Trends on Travel Demand in Georgia*. Final Report, Georgia Department of Transportation Research Project 16-31, available from the authors and at http://g92018.eos-intl.net/G92018/OPAC/Index.aspx.

Konduri, K. C., Astroza, S., Sana, B., Pendyala, R. M., & Jara-Díaz, S. R. (2011). Joint Analysis of Time Use and Consumer Expenditure Data: Examination of Two Approaches to Deriving Values of Time. *Transportation Research Record, 2231*(1), 53-60. doi:10.3141/2231-07

Kressner, J. D., Macfarlane, G. S., Huntsinger, L., & Donnelly, R. (2016). *Using passive data to build an agile tour-based model: a case study in Asheville*. Paper presented at the International Conference on Innovations in Travel Modeling, Denver, CO. Retrieved from: https://pdfs.semanticscholar.org/27ae/db2df8e8709ece22d2042aea75d403df9285.pdf

Kressner, J. D. (2017). *Synthetic Household Travel Data Using Consumer and Mobile Phone Data* (Report No. 184). Washington D.C.: National Cooperative Highway Research Program (NCHRP) Innovations Deserving Exploratory Analysis (IDEA) Program, Transportation Research Board. Retrieved from: http://www.trb.org/Research/Blurbs/176216.aspx

Kressner, J. D., & Garrow, L. A. (2012). Lifestyle segmentation variables as predictors of home-based trips for Atlanta, Georgia, airport. *Transportation Research Record: Journal of the Transportation Research Board, 2266*, 20-30.

Kressner, J. D., & Garrow, L. A. (2014). Using third-party data for travel demand modeling: comparison of targeted marketing, census, and household travel survey data. *Transportation Research Record: Journal of the Transportation Research Board, 2442*, 8–19.

Kressner, J. D., Carragher, M. F., & Watkins, K. E. (2014). *A household-level pairwise comparison of targeted marketing data and self-reported survey data*. Paper presented at the 93rd Annual Meeting of the Transportation Research Board, Washington D.C. Retrieved from: https://www.researchgate.net/publication/341909525_A_Household-

Level_Pairwise_Comparison_of_Targeted_Marketing_Data_and_Self-Reported_Survey_Data

Kuppam, A. R., Pendyala, R. M., & Rahman, S. (1999). Analysis of the Role of Traveler Attitudes and Perceptions in Explaining Mode-Choice Behavior. *Transportation Research Record, 1676*(1), 68-76. doi:10.3141/1676-0

Kupper, N., & Denollet, J. (2012). Social anxiety in the general population: Introducing abbreviated versions of SIAS and SPS. *Journal of Affective Disorders, 136*(1), 90-98. doi: 10.1016/j.jad.2011.08.014

Laney, D. (2001). 3D data management: controlling data volume, velocity, and variety. *META Group Research Note 6*.

Lavalle, S., Lesser, E., Shockley, R., S. Hopkins, M., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT Sloan Management Review, 52*(2), 21-31.

Lohr, S. L., & Raghunathan, T. E. (2017). Combining survey data with other data sources. *Statistical Science, 32*(2), 293-312. doi:10.1214/16-STS584

Lorenz-Spreen, P., Mønsted, B. M., Hövel, P., & Lehmann, S. (2019). Accelerating dynamics of collective attention. *Nature Communications*, 10(1), 1759. doi:10.1038/s41467-019-09311-w

Lovelace, R., Birkin, M., Cross, P., & Clarke, M. (2016). From Big Noise to Big Data: Toward the Verification of Large Data sets for Understanding Regional Retail Flows. *Geographical Analysis, 48*(1), 59-81. doi: 10.1111/gean.12081

Lukoianova, T., & Rubin, V. L. (2014). Veracity roadmap: Is big data objective, truthful and credible? *Advances in Classification Research Online; 24th ASIS SIG/CR Classification Research Workshop*. Retrieved from https://journals.lib.washington.edu/index.php/acro/article/view/14671/12311

Ma, J., Li, H., Yuan, F., & Bauer, T. (2013). Deriving operational origin-destination matrices from large scale mobile phone data. *International Journal of Transportation Science and Technology, 2*(3), 183-204. doi: 10.1260/2046-0430.2.3.183

Macfarlane, G. S. (2014). *Using Big Data to Model Travel Behavior: Applications to Vehicle Ownership and Willingness-to-Pay for Transit Accessibility.* [Doctoral dissertation, Georgia Institute of Technology]. SMARTech Repository. https://smartech.gatech.edu/handle/1853/51804

Macfarlane, G. S., Garrow, L. A., & Mokhtarian, P. L. (2015). The influences of past and present residential locations on vehicle ownership decisions. *Transportation*

Research Part A: Policy and Practice, 74, 186-200. doi: 10.1016/j.tra.2015.01.005

Macfarlane, G. S., Garrow, L. A., & Moreno-Cruz, J. (2015). Do Atlanta residents value MARTA? Selecting an autoregressive model to recover willingness to pay. *Transportation Research Part A: Policy and Practice, 78*, 214-230. doi: 10.1016/j.tra.2015.05.010

Maio, G. R., & Haddock, G. (2009). *The Psychology of Attitudes and Attitude Change*. Thousand Oaks, California: Sage Publications Ltd.

Malokin, A. (2019). *Pathways to Improving Traditional Travel Bheavior Models with Travel-based Multitasking and Attitudinal Data*. [Doctoral dissertation, Georgia Institute of Technology]. SMARTech Repository. https://smartech.gatech.edu/handle/1853/4760/browse?value=Malokin%2C+Aliaksandr&type=author

Marsh, H. W., Huppert, F. A., Donald, J. N., Horwood, M. S., & Sahdra, B. K. (2020). The well-being profile (WB-Pro): Creating a theoretically based multidimensional measure of well-being to advance theory, research, policy, and practice. *Psychological Assessment, 32*(3), 294-313. doi:10.1037/pas0000787

Minaee, S. (2019). 20 *Popular Machine Learning Metrics, Part 1: Classification & Regression Evaluation Metrics*. Retrieved September 9, 2020 from: https://towardsdatascience.com/20-popular-machine-learning-metrics-part-1-classification-regression-evaluation-metrics-1ca3e282a2ce

Mokhtarian, P., & Salomon, I. (1997). Modeling the desire to telecommute: The importance of attitudinal factors in behavioral models. *Transportation Research Part A: Policy and Practice, 31*(1), 35-50.

Müller, K., & Axhausen, K. W. (2014). Using survey calibration and statistical matching to reweight and distribute activity schedules. *Transportation Research Record*, 2429, 157-167. doi:10.3141/2429-17

Murphy, K. M., & Topel, R. H. (2002). Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics*, 20(1), 88-97. doi: 10.1198/073500102753410417

National Academy of Sciences, National Academy of Engineering, & Institute of Medicine. (2007). *Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future*. Washington, DC: The National Academies Press. Retrieved from: https://www.nap.edu/catalog/11463/rising-above-the-gathering-storm-energizing-and-employing-america-for

National Research Council. (2013). *Nonresponse in Social Science Surveys : A Research Agenda*. Washington, DC: The National Academies Press. Retrieved from: https://www.nap.edu/catalog/18293/nonresponse-in-social-science-surveys-a-research-agenda

Newcombe, H. B., Kennedy J. M., Axford S. J., and James A. P. (1959). Automatic linkage of vital records. *Science*, *130*(3381), 954-959.

Nicolaisen, M. S., & Driscoll, P. A. (2014). Ex-post evaluations of demand forecast accuracy: A literature review. *Transport Reviews, 34*(4), 540-557. doi:10.1080/01441647.2014.926428

Okner, B. (1974). Data matching and merging: an overview. *Annals of Economic and Social Measurement*, *3*(2), 347-352.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345-1359. doi: 10.1109/tkde.2009.191

Parady, G., Ory, D., & Walker, J. (2021). The overreliance on statistical goodness-of-fit and under-reliance on model validation in discrete choice models: A review of validation practices in the transportation academic literature. *Journal of Choice Modelling, 38*, 100257. doi: 10.1016/j.jocm.2020.100257

Parthasarathi, P., & Levinson, D. (2010). Post-construction evaluation of traffic forecast accuracy. *Transport Policy, 17*(6), 428-443. doi:10.1016/j.tranpol.2010.04.010

Pawlak, J., Polak, J., & Sivakumar, A. (2013). *An imputation approach to the fusion of travel diary and lifestyle data: Application the analysis of the interaction of ICT and physical mobility.* Paper presented at the the New Techniques and Technologies for Statistics conference, Brussels, Belgium.

PTV NuStats. (2011). *Regional Travel Survey: Final Report*. Atlanta, Georgia. Retrieved from: https://cdn.atlantaregional.org/wp-content/uploads/tp-2011regionaltravelsurvey-030712.pdf

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41(1), 203-212. doi:10.1016/j.jrp.2006.02.001

Ramsey, K., & Bell, A. (2014). *Smart Location Database: Version 2.0 User Guide*. Retrieved January 21, 2021 from: https://www.epa.gov/sites/production/files/2014-03/documents/sld_userguide.pdf

Rubin, D. B. (1987). *Multiple Imputation for Non-response in Surveys*. Chichester, England: John Wiley & Sons.

Ruiz, T., Mars, L., Arroyo, R., & Serna, A. (2016). Social Networks, Big Data and Transport Planning. *Transportation Research Procedia, 18*, 446-452. doi: 10.1016/j.trpro.2017.01.122

Sahdra, B. K., Ciarrochi, J., Parker, P., & Scrucca, L. (2016). Using Genetic Algorithms in a Large Nationally Representative American Sample to Abbreviate the Multidimensional Experiential Avoidance Questionnaire. *Frontiers in Psychology, 7*(189). doi:10.3389/fpsyg.2016.00189

Salomon, I., & Ben-Akiva, M. (1983). The Use of the Life-Style Concept in Travel Demand Models. *Environment and Planning A: Economy and Space, 15*(5), 623-638. doi: 10.1068/a150623

Sandy, C. J., Gosling, S. D., & Koelkebeck, T. (2014). Psychometric Comparison of Automated Versus Rational Methods of Scale Abbreviation. *Journal of Individual Differences, 35*(4), 221-235. doi: 10.1027/1614-0001/a000144

Saporta, G. (2002). Data fusion and data grafting. *Computational Statistics and Data Analysis*, *38*(4), 465-473. doi: 10.1016/S0167-9473(01)00072-X

Shaw, F. A., Malokin, A., Mokhtarian, P. L., & Circella, G. (2019). Who doesn't mind waiting? Examining the relationships between waiting attitudes and person- and travel-related attributes. *Transportation*. doi: 10.1007/s11116-019-10054-2

Shaw, F. A., Wang, X., Mokhtarian, P. & Watkins, K. (paper under review, available upon request from authors). Supplementing transportation data sources with targeted marketing data: Applications, integration, and validation.

Solon, G., Haider, S. J., & Woolridge, J. M. (2015). What are we weighting for? *Journal of Human Resources*, *50*(2), 301-316. doi:10.3368/jhr.50.2.301

Sivakumar, A., & Polak, J. (2013). An exploration of data pooling techniques: Modeling activity participation and household technology holdings. Paper presented at the 92nd Annual Meeting of the Transportation Research Board, Washington D.C.

Sivakumar, A., & Polak, J. (2009). Modelling the endogeneity in activity participation and technology holdings: An exploration of data pooling techniques. Paper presented at the International Choice Modelling Conference, Harrogate, England.

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of big data challenges and analytical methods. *Journal of Business Research, 70*, 263-286. doi:10.1016/j.jbusres.2016.08.001

Stevens, J. P. (2009). *Applied Multivariate Statistics for the Social Sciences*. New York, NY: Taylor & Francis Group, LLC.

Tobias, E., Ralf, M., & Christian, B. (2013). On the impact of response patterns on survey estimates from access panels. *Survey Research Methods, 7*(2)**,** 91-101. doi:10.18148/srm/2013.v7i2.5036

Toole, J. L., Colak, S., Sturt, B., Alexander, L. P., Evsukoff, A., & González, M. C. (2015). The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies, 58*, 162-177. doi:10.1016/j.trc.2015.04.022

Tsamardinos, I., Triantafillou, S., & Lagani, V. (2012). Towards integrative causal analysis of heterogenous data sets and studies. *Journal of Machine Learning Research*, *13*(1), 1097-1157.

U.S. Census Bureau. (2020). *2013-2017 American Community Survey, 5-year estimates*. Retrieved February, 2020 from: https://www.census.gov/

Van Acker, V., Goodwin, P., & Witlox, F. (2016). Key research themes on travel behavior, lifestyle, and sustainable urban mobility. *International Journal of Sustainable Transportation, 10*(1), 25-32. doi:10.1080/15568318.2013.821003

van der Putten, P. , Kok, J. N., & Gupta, A. (2002). Data Fusion through Statistical Matching. MIT Sloan School of Management, Working paper 4342-02. Retrieved from http://liacs.leidenuniv.nl/~puttenpwhvander/library/2002fusionsloan.pdf

Voulgaris, C. T. (2019). Crystal balls and black boxes: What makes a good forecast? *Journal of Planning Literature, 34*(3), 286-299. doi:10.1177/0885412219838495

Walker, J., Vij, A., & Brathwaite, T. (2019). Choice modelling in an age of machine learning. Keynote presentation at the International Choice Modelling Conference, Kobe, Japan.

Wang, F., & Chen, C. (2018). On data processing required to derive mobility patterns from passively-generated mobile phone data. *Transportation Research Part C: Emerging Technologies, 87*, 58-74. https://doi.org/10.1016/j.trc.2017.12.003

Wang, F., & Ross, C. L. (2018). Machine learning travel mode choices: Comparing the performance of an extreme gradient boosting model with a multinomial logit model. *Transportation Research Record*, 2672(47), 35-45. doi:10.1177/0361198118773556

Wang, X., Shaw, F. A., Mokhtarian, P. L., Circella, G., & Watkins, K. E. (paper under review, available upon request from authors). Combining disparate surveys across time to study satisfaction with life: the effects of student context, sampling method, and transport attributes.

Wang, X., Shaw, F. A., Mokhtarian, P. L., & Watkins, K. E. (paper in preparation, available upon request from authors). Respondent profiles of willingness to respond in consecutive travel surveys: An investigation based on the U.S. National Household Travel Survey.

Wang, Z., He, S. Y., & Leung, Y. (2018). Applying mobile phone data to travel behaviour research: A literature review. *Travel Behaviour and Society, 11*, 141-155. doi: 10.1016/j.tbs.2017.02.005

Wassenaar, A., de Reus, J., Donders, A. R. T., Schoonhoven, L., Cremer, O. L., de Lange, D. W., . . . van den Boogaard, M. (2018). Development and Validation of an Abbreviated Questionnaire to Easily Measure Cognitive Failure in ICU Survivors: A Multicenter Study. *Critical Care Medicine, 46*(1), 79-84. doi: 10.1097/CCM.0000000000002806

Welch, T. F., & Widita, A. (2019). Big data in public transportation: a review of sources and methods. *Transport Reviews*, 1-24. doi:10.1080/01441647.2019.1616849

Welde, M., & Odeck, J. (2011). Do planners get it right? The accuracy of travel demand forecasting in Norway. *European Journal of Transport and Infrastructure Research, 11*(1). doi:10.18757/ejtir.2011.11.1.2913

Winkler, W. E. (1999). *The state of record linkage and current research problems*. Technical Report, Statistical Research Division. U.S. Census Bureau.

Yarkoni, T. (2010). The abbreviation of personality, or how to measure 200 personality scales with 200 items. *Journal of Research in Personality, 44*(2), 180-198. doi: 10.1016/j.jrp.2010.01.002

Yiu, T. (2019a). *The Curse of Dimensionality: Why High Dimensional Data Can be So Troublesome*. Retrieved September 9, 2020 from: https://towardsdatascience.com/the-curse-of-dimensionality-50dc6e49aa1e

Yiu, T. (2019b). *Understanding PCA (Principal Components Analysis)*. Retrieved September 9, 2020 from: https://towardsdatascience.com/understanding-pca-fae3e243731d

Zhao, X., Yan, X., Yu, A., & Van Hentenryck, P. (2020). Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. *Travel Behaviour and Society, 20*, 22-35. doi:10.1016/j.tbs.2020.02.003

Zheng, Y. (2015). Methodologies for cross-domain fusion: An overview. *IEEE Transactions on Big Data*, *1*(1), 16-34. doi:10.1109/tbdata.2015.2465959